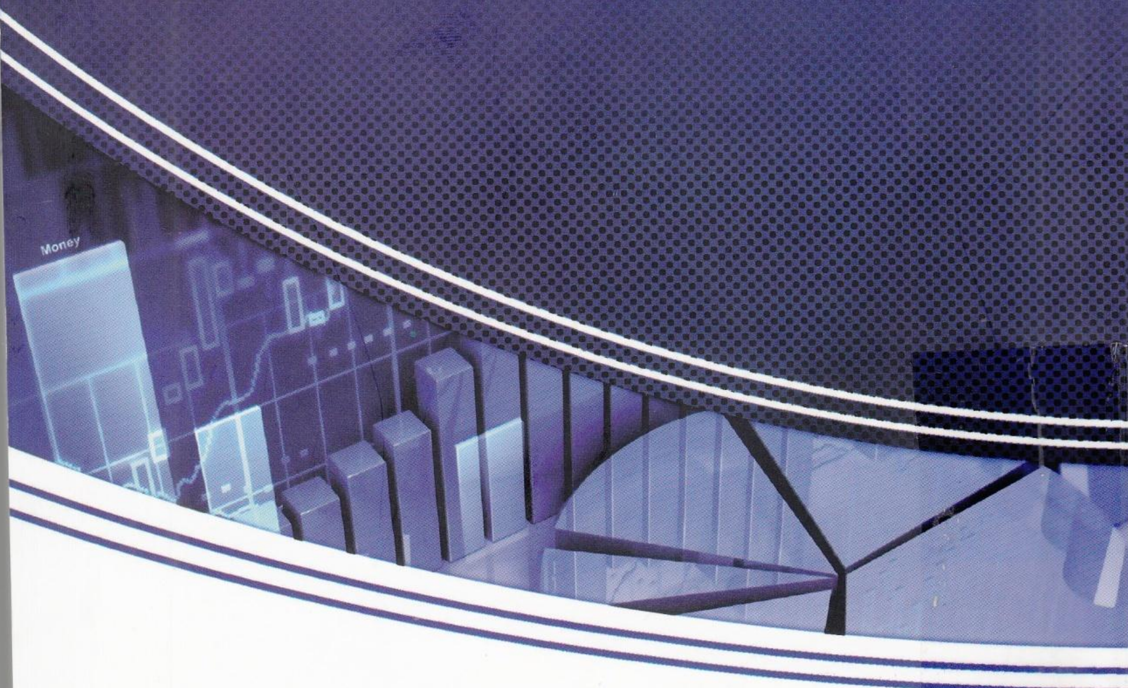
به همراه CD

مقدماتی

اقتصاد سنجی (جلد ۱)

همراه با کاربرد Eviews 8 & Stata 12



دکتر علی سوری

فهرست مطالب

عنوان | صفحه

جلد اول

مقدمه ۱
فصل اول: مروری بر مفاهیم پایه‌ای آمار و نرم‌افزارهای Eviews و Stata ۵
۱-۱ مقدمه ۵
۱-۲ نمونه تصادفی ۵
۱-۳ توزیع مشترک متغیرهای نمونه یا توزیع مشترک نمونه تصادفی ۶
۱-۴ تابعی از متغیرهای نمونه (آماره) ۸
۱-۵ تخمین ۹
۱-۵-۱ تخمین و تخمین‌زننده ۹
۱-۵-۲ روش‌های تخمین ۱۱
روش گشتاورها ۱۱
روش حداکثر درستنمایی ۱۲
روش حداقل مربعات معمولی ۱۴
۱-۵-۳ خواص تخمین‌زننده‌ها ۱۴
بدون تورش یا ناریب بودن ۱۵
کارایی (حداقل واریانس) ۱۶
سازگاری ۱۸
کفایت ۱۹
۱-۵-۴ تخمین فاصله‌ای ۲۱
۱-۶ آزمون فرضیه ۲۳
خطای نوع اول و نوع دوم ۲۴
۱-۷ مروری بر توزیع‌های مهم ۲۵
۱-۷-۱ توزیع نرمال استاندارد ۲۵

۲-۷-۱ توزیع کای‌دو ۲۵
۳-۷-۱ توزیع t ۲۸
۴-۷-۱ توزیع F ۲۹
۵-۷-۱ حالت‌های خاص توزیع‌های χ^2 ، t و F ۳۰
۸-۱ انواع داده‌ها ۳۱
۹-۱ آشنایی مقدماتی با نرم‌افزار Eviews ۳۲
مسائل ۳۹
ضمیمه فصل اول: مروری بر نرم افزار Stata ۴۲

فصل دوم: رگرسیون ساده ۵۳
۱-۲ مقدمه ۵۳
۲-۲ امید ریاضی شرطی و رگرسیون ۵۳
رگرسیون نرمال ۵۵
۳-۲ جملهٔ خطا و معادله رگرسیون ۵۷
۴-۲ فروض معادله رگرسیون ۵۸
۵-۲ تحلیل رگرسیون تجربی ۶۲
۶-۲ تخمین معادله رگرسیون ۶۳
۷-۲ خواص تخمین‌زننده‌های OLS ۶۶
۱-۷-۲ تخمین‌زنندهٔ خطی ۶۷
۲-۷-۲ بدون تورش بودن ۶۷
۳-۷-۲ سازگاری ۶۸
۴-۷-۲ حداقل واریانس ۶۸
۵-۷-۲ توزیع تخمین‌زننده‌های OLS ۶۹
۶-۷-۲ همبستگی $\hat{\alpha}$ و $\hat{\beta}$ ۶۹
۸-۲ تغییرات کل، تغییرات توضیح داده شده و تغییرات توضیح داده نشده ۷۱
۹-۲ ضریب تعیین (R^2) ۷۲
۱۰-۲ میانگین خطای تخمین یا انحراف معیار رگرسیون($\hat{\sigma}$) ۷۳
۱۱-۲ آزمون معنادار بودن ضرایب رگرسیون ۷۵
۱۲-۲ تحلیل واریانس (آزمون معنادار بودن رگرسیون) ۷۷
۱۳-۲ جمع‌بندی و تحلیل نتایج رگرسیون ۸۰
۱۴-۲ پیش‌بینی و فاصله اطمینان پیش‌بینی ۸۳
۱۵-۲ رگرسیون‌های غیرخطی ۸۷
۱-۱۵-۲ روابط معکوس ۸۷

۲-۱۵-۲ معادلات تمام لگاریتمی (log-log) ۸۸
۳-۱۵-۲ توابع نمایی ۹۰
۴-۱۵-۲ رگرسیون انحراف از میانگین ۹۱
۵-۱۵-۲ رگرسیون با متغیرهای استاندارد شده ۹۲
۷-۱۵-۲ برآورد معادله روند ۹۳
۸-۱۵-۲ برآورد معادله نرخ رشد ۹۴
۱۶-۲ تحلیل همبستگی ۹۶
۱-۱۶-۲ کوواریانس ۹۶
۲-۱۶-۲ ضریب همبستگی ۹۹
مسائل ۱۰۴
ضمیمه فصل دوم: برآورد رگرسیون ساده در Stata ۱۱۰

فصل سوم: رگرسیون چندمتغیره ۱۱۹
۱-۳ مقدمه ۱۱۹
۲-۳ رگرسیون دو متغیره: مفاهیم و فروض ۱۱۹
۳-۳ تخمین ضرایب رگرسیون دو متغیره و خواص آنها ۱۲۰
۴-۳ ضریب تعیین و خطای معادلهٔ رگرسیون ۱۲۳
۵-۳ معیارهای اطلاعات ۱۲۵
۶-۳ آزمون معنی‌دار بودن ضرایب معادله رگرسیون ۱۲۶
۷-۳ آزمون معنی‌دار بودن معادله رگرسیون (تحلیل واریانس) ۱۲۷
۸-۳ تحلیل نتایج رگرسیون دو متغیره ۱۲۸
۹-۳ تحلیل همبستگی چندمتغیره ۱۳۰
۱۰-۳ مدل Kمتغیره ۱۳۳
۱۱-۳ رگرسیون مقید و تخمین‌زننده‌های مقید ۱۴۰
۱۲-۳ آزمون محدودیت‌ها ۱۴۳
۱۳-۳ پیش‌بینی با رگرسیون چندمتغیره ۱۴۸
مسائل ۱۵۰
ضمیمه فصل سوم: برآورد رگرسیون چند متغیره در Stata ۱۵۵

فصل چهارم: نقض فروض کلاسیک ۱۵۹
۱-۴ مقدمه ۱۵۹
۲-۴ فرض اول: صفر بودن میانگین خطا ۱۵۹
۳-۴ فرض دوم: واریانس همسانی ۱۶۳

۱-۳-۴ ماهیت واریانس ناهمسانی ۱۶۳
۲-۳-۴ پیامدهای واریانس ناهمسانی ۱۶۵
۳-۳-۴ برآورد واریانس مستحکم وایت ۱۶۵
۴-۳-۴ آزمون‌های تشخیص واریانس ناهمسانی ۱۶۹
الف) آزمون بارتلت ۱۶۹
ب) آزمون گلدفلد- کوانت ۱۷۱
ج) آزمون گلجسر ۱۷۲
د) آزمون همبستگی رتبه‌ای اسپیرمن ۱۷۲
هـ) آزمون بروش- پاگان-گادفری ۱۷۳
و) آزمون هاروی ۱۷۴
ز) آزمون وایت ۱۷۵
۵-۳-۴ تخمین ضرایب با وجود واریانس ناهمسانی ۱۷۹
۴-۴ فرض سوم: عدم خودهمبستگی ۱۸۵
۱-۴-۴ مفهوم وقفه ۱۸۵
۱-۴-۴ پیامدهای خودهمبستگی ۱۸۵
۲-۴-۴ روش‌های نموداری جهت تشخیص خودهمبستگی ۱۸۶
۳-۴-۴ آزمون دوربین- واتسون ۱۹۰
۴-۴-۴ آزمون بروش-گادفری ۱۹۴
۵-۴-۴ پیامد نادیده گرفتن خودهمبستگی ۱۹۶
۶-۴-۴ تخمین ضرایب در حالت خودهمبستگی ۱۹۷
۷-۴-۴ خودهمبستگی و مدل‌های پویا ۱۹۹
۸-۴-۴ خودهمبستگی و فرایندهای خودرگرسیون (AR) و میانگین متحرک ۲۰۲
۵-۴ فرض ۴: غیر تصادفی بودن متغیرهای توضیحی ۲۰۳
۶-۴ فرض ۵: فرض نرمال بودن u_t ۲۰۴
مسائل ۲۰۸
ضمیمه فصل چهارم: آزمون نقض فروض در Stata ۲۱۲

فصل پنجم: متغیرهای مجازی ۲۱۷
۱-۵ مقدمه ۲۱۷
۲-۵ متغیر مجازی ۲۱۷
۳-۵ رگرسیون روی متغیر مجازی ۲۱۸
۴-۵ دام متغیرهای مجازی ۲۲۰
۴-۵ رگرسیون روی متغیر مجازی همراه با متغیر توضیحی ۲۲۱

۵-۵ متغیر مجازی و تغییر شیب ۲۲۱
۵-۶ وجود دو عامل کیفی ۲۲۳
۵-۷ تأثیر متقابل دو عامل کیفی ۲۲۴
۵-۸ متغیرهای مجازی و تغییر ساختار ۲۲۴
۵-۹ متغیرهای مجازی و مشاهدات دور افتاده ۲۲۵
۵-۱۰ متغیرهای مجازی و تغییرات فصلی ۲۲۸
۵-۱۱ توابع لگاریتمی و متغیر مجازی ۲۲۹
۵-۱۲ قیمت‌گذاری عوامل کیفی ۲۳۰
مسائل ۲۳۲
ضمیمه فصل پنجم: متغیرهای مجازی در Stata ۲۳۴

فصل ششم: آزمون‌های خاص: تصریح مدل، تغییر ساختاری و علیّت ۲۳۷
۶-۱ مقدمه ۲۳۷
۶-۲ همخطی ۲۳۷
۶-۲-۱ مفهوم همخطی ۲۳۷
۶-۲-۲ مشکلات ناشی از همخطی ۲۳۸
۶-۲-۳ شناسایی همخطی ۲۳۹
۶-۲-۴ راه‌های کاهش همخطی ۲۴۲
۶-۳ فرم تابعی نادرست ۲۴۴
۶-۴ حذف متغیرهای مهم ۲۴۸
۶-۵ ورود متغیرهای نامربوط ۲۵۳
۶-۶ آزمون‌های ثبات ضرایب ۲۵۶
۶-۶-۱ آزمون نقطه شکست چاو ۲۵۶
۶-۶-۲ آزمون پیش‌بینی چاو ۲۵۸
۶-۶-۳ آزمون‌های بازگشتی ۲۶۰
۶-۶-۴ پیش‌بینی یک‌قدمی ۲۶۴
۶-۶-۵ آزمون مجموع تجمعی خطاهای بازگشتی (CUSUM) ۲۶۶
۶-۶-۶ آزمون مجموع مجذور تجمعی خطاهای بازگشتی (CUSUMQ) ۲۶۸
۶-۶-۷ متغیرهای مجازی ۲۶۹
۶-۷ خطای اندازه‌گیری متغیرها ۲۷۰
۶-۹ علّیت ۲۷۳
مسائل ۲۷۵
ضمیمه فصل ششم: آزمون‌های خاص در Stata ۲۷۸

جلد دوم

فصل هفتم: روش حداکثر درستنمایی ۲۸۳

۷-۱ مقدمه ۲۸۳

۷-۲ تابع درستنمایی ۲۸۳

۷-۳ تخمین‌زنندهٔ حداکثر درستنمایی ۲۸۵

۷-۴ روش حداکثر درستنمایی در رگرسیون ساده ۲۸۶

۷-۵ روش حداکثر درستنمایی در رگرسیون چند متغیره ۲۸۸

۷-۶ روش حداکثر درستنمایی در حالت واریانس ناهمسانی (روش GLS) ۲۹۰

۷-۶ شبه R^2 ۲۹۲

۷-۷ آزمون معنی‌دار بودن رگرسیون ۲۹۳

۷-۸ تخمین‌زنندهٔ حداکثر درستنمایی مقید ۲۹۴

۷-۹ شرط مرتبه دوم ۲۹۵

۷-۱۰ ماتریس امتیاز ۲۹۶

۷-۱۱ ماتریس اطلاعات ۲۹۹

۷-۱۲ رابطه ماتریس امتیاز و ماتریس اطلاعات ۳۰۰

۷-۱۳ نامساوی کرامر- رائو ۳۰۲

۷-۱۴ آزمون نسبت درستنمایی (LR) ۳۰۴

۷-۱۵ آزمون ضریب لاگرانژ ۳۰۶

۷-۱۶ آزمون والد ۳۰۸

۷-۱۷ مقایسه آماره‌های نسبت درستنمایی، والد و ضریب لاگرانژ ۳۱۰

مسائل ۳۱۳

فصل هشتم: متغیرهای ابزاری(IV) و حداقل مربعات دومرحله‌ای(2SLS) ۳۱۵

۸-۱ مقدمه ۳۱۵

۸-۲ متغیرهای ابزاری (IV) در رگرسیون ساده ۳۱۵

۸-۳ واریانس تخمین‌زنندهٔ IV ۳۱۹

۸-۴ ضریب تعیین (R^2) ۳۲۱

۸-۵ متغیرهای ابزاری در رگرسیون چند متغیره ۳۲۲

۸-۶ تخمین‌زنندهٔ IV در حالت عمومی ۳۲۴

۸-۷ تخمین‌زننده حداقل مربعات دومرحله‌ای (2SLS) ۳۲۷

۸-۸ آماره F در روش متغیرهای ابزاری ۳۲۸

۸-۹ آزمون سارگان برای بررسی اعتبار متغیرهای ابزاری ۳۲۹

۸-۱۰ آزمون هاسمن ۳۳۰

مسائل ۳۳۳

ضمیمه فصل هشتم: آزمون‌های خاص در Stata ۳۳۴

فصل نهم: مدل‌های غیرخطی ۳۳۷

۱-۹ مقدمه ۳۳۷

۲-۹ رگرسیون خطی ۳۳۷

۳-۹ رگرسیون‌های غیرخطی قابل تبدیل به خطی ۳۳۸

۴-۹ رگرسیون غیرخطی ۳۳۸

۵-۹ فروض مدل رگرسیون غیرخطی ۳۴۰

۶-۹ تخمین‌زنندهٔ حداقل مربعات غیرخطی ۳۴۱

۱-۶-۹ الگوریتم گاوس- نیوتن ۳۴۱

۲-۶-۹ الگوریتم نیوتن- رافسون ۳۵۱

۳-۶-۹ مقایسه الگوریتم‌های گاوس- نیوتن و نیوتن- رافسون ۳۵۱

۷-۹ روش حداکثر درستنمایی ۳۵۵

۸-۹ آزمون محدودیت‌ها در مدل‌های غیرخطی ۳۶۳

مسائل ۳۶۸

ضمیمه فصل نهم: رگرسیون غیر خطی در Stata ۳۶۹

فصل ۱۰: مدل‌های با وقفهٔ توزیعی ۳۷۱

۱-۱۰ مقدمه ۳۷۱

۲-۱۰ اثرات تأخیری ۳۷۱

۳- ۱۰ تخمین مدل‌های با وقفهٔ توزیعی ۳۷۲

۴-۱۰ الگوی وقفهٔ خطی ۳۷۶

۵-۱۰ الگوی وقفهٔ V معکوس ۳۷۸

۶-۱۰ الگوی وقفهٔ چندجمله‌ای: روش آلمون ۳۸۰

۷-۱۰ مدل‌های با وقفهٔ نامحدود: تبدیل کویک ۳۸۸

۸-۱۰ الگوی وقفهٔ پاسکال ۳۹۲

۹-۱۰ مبانی نظری مدل‌های با وقفهٔ توزیعی ۳۹۴

۱۰-۱۰ مدل‌های خودرگرسیون با وقفهٔ توزیعی (ARDL) ۳۹۷

مسائل ۴۰۴

ضمیمه فصل دهم: برآورد مدل‌های با وقفهٔ توزیعی در Stata ۴۰۶

فصل یازدهم: سری‌های زمانی یک متغیره ۴۰۷
۱-۱۱ مقدمه ۴۰۷
۲-۱۱ برخی مفاهیم سری‌های زمانی ۴۰۸
۱-۲-۱۱ فرایند اکیداً مانا ۴۰۸
۲-۲-۱۱ فرایند مانای ضعیف ۴۰۸
۳-۲-۱۱ فرایند تصادفی محض ۴۰۹
۴-۲-۱۱ آزمون معنادار بودن ضرایب خودهمبستگی ۴۱۰
۳-۱۱ فرایند میانگین متحرک ۴۱۴
۴-۱۱ فرایند خودرگرسیون ۴۲۱
۵-۱۱ قضیه تجزیه ولد ۴۳۴
۶-۱۱ ضرایب خودهمبستگی جزئی ۴۳۵
۷-۱۱ معکوس‌پذیری MA(q) ۴۳۶
۸-۱۱ فرایند ARMA ۴۳۹
۹-۱۱ مدل سازی ARMA : روش باکس- جنکینز ۴۴۰
۱۰-۱۱ استفاده از معیارهای اطلاعات برای انتخاب مدل ARMA ۴۴۴
۱۱-۱۱ مدل‌های ARIMA ۴۴۵
۱۲-۱۱ پیش‌بینی با استفاده از مدل‌های سری زمانی ۴۴۶
۱-۱۲-۱۱ معیارهای ارزیابی پیش‌بینی ۴۴۶
۲-۱۲-۱۱ پیش‌بینی با مدل MA ۴۴۹
۳-۱۲-۱۱ پیش‌بینی با مدل AR ۴۵۰
۴-۱۲-۱۱ پیش‌بینی ایستا و پویا ۴۵۱
مسائل ۴۵۵
ضمیمه فصل یازدهم: مدل‌های ARIMA در Stata ۴۵۷

فصل دوازدهم: مانایی، ریشه واحد و هم‌انباشتگی ۴۶۱
۱-۱۲ مقدمه ۴۶۱
۲-۱۲ مانایی ۴۶۱
۳-۱۲ مانایی ضعیف ۴۶۲
۴-۱۲ سری‌های زمانی مانا ۴۶۳
۵-۱۲ روند قطعی ۴۶۵
۶-۱۲ روند تصادفی ۴۶۶
۷-۱۲ ترکیب روند قطعی و روند تصادفی ۴۶۹
۸-۱۲ روندزدایی ۴۷۱

۹-۱۲ آزمون ریشه واحد ۴۷۲
۱۰-۱۲ آزمون ریشه واحد و تغییر ساختاری ۴۸۰
۱۱-۱۲ هم‌انباشتگی ۴۸۵
۱۲-۱۲ آزمون هم‌انباشتگی ۴۸۹
۱۳-۱۲ مدل‌های تصحیح خطا (ECM) ۴۹۲
۱۴-۱۲ تخمین مدل تصحیح خطا ۴۹۴
مسائل ۴۹۵
ضمیمه فصل دوازدهم: آرمون‌های مانایی در Stata ۴۹۸

فصل سیزدهم: سری‌های زمانی فصلی ۵۰۱
۱-۱۳ مقدمه ۵۰۱
۲-۱۳ الگوی فصلی قطعی ۵۰۲
۱-۲-۱۳ متغیرهای مجازی فصلی ۵۰۲
۲-۲-۱۳ الگوهای مثلثاتی ۵۰۴
۳-۱۳ مدل‌های ARMA فصلی (SARMA) ۵۰۹
۴-۱۳ مدل ARIMA فصلی (SARIMA) ۵۱۳
۵-۱۳ ضرایب خودهمبستگی سری‌های زمانی فصلی ۵۱۶
۶-۱۳ ریشه واحد فصلی ۵۱۹
۷-۱۳ آزمون ریشه واحد فصلی ۵۲۳
مسائل ۵۳۰
ضمیمه فصل سیزدهم: ریشه واحد فصلی در Stata ۵۳۱

فصل چهاردهم: مدل‌های تغییرپذیری ۵۳۵
۱-۱۴ مقدمه ۵۳۵
۲-۱۴ واریانس ناهمسانی ۵۳۵
۳-۱۴ تغییرپذیری و نااطمینانی ۵۳۹
۴-۱۴- مدل ARCH ۵۴۰
۵-۱۴ آزمون ARCH ۵۴۲
۶-۱۴ محدودیت‌های مدل ARCH ۵۴۵
۷-۱۴- مدل ARCH تعمیم‌یافته (GARCH) ۵۴۵
۸-۱۴ تخمین مدل‌های ARCH و GARCH ۵۴۶
۹-۱۴ مدل GARCH نامتقارن ۵۵۲
۱-۹-۱۴ مدل GJR ۵۵۳

۱۴-۹-۲ مدل TARCH	۵۵۴
۱۴-۹-۳ مدل EGARCH	۵۵۶
۱۴-۹-۴ آزمون عدم‌تقارن	۵۵۸
۱۴-۹-۵ منحنی تأثیر خبر	۵۶۱
۱۴-۱۰ وارد نمودن GARCH در معادله میانگین شرطی (ARCH-M)	۵۶۲
۱۴-۱۱ GARCH با اجزاء موقتی و دائمی	۵۶۴
۱۴-۱۲ پیش‌بینی با مدل‌های GARCH	۵۶۵
۱۴-۱۳ مدل‌های GARCH چندمتغیره (MGARCH)	۵۶۹
مسائل	۵۷۷
ضمیمه فصل چهاردهم: مدل‌هایGARCH در Stata	۵۸۰
فصل پانزدهم: مدل‌های تغییر جهت	۵۸۷
۱۵-۱ مقدمه	۵۸۷
۱۵-۲ تغییرات فصلی	۵۸۹
۱۵-۳توابع خطی قطعه‌ای	۵۹۱
۱۵-۴ مدل‌های تغییر جهت مارکف	۵۹۳
۱۵-۴-۱ مبانی مدل‌های تغییر جهت مارکف	۵۹۴
۱۵-۴-۲ کاربردی از مدل تغییر جهت مارکف	۶۰۱
۱۵-۵ مدل‌های خودرگرسیون آستانه	۶۰۴
مسائل	۶۰۹
فصل شانزدهم: معادلات به‌ظاهر نامرتبط (SUR)	۶۱۱
۱۶-۱ مقدمه	۶۱۱
۱۶-۲ الگوی سادهٔ معادلات به‌ظاهر نامرتبط	۶۱۱
۱۶-۳ تخمین‌زننده‌های OLS در معادلات به‌ظاهر نامرتبط	۶۱۲
۱۶-۴ تخمین‌زننده‌های GLS در معادلات به‌ظاهر نامرتبط	۶۱۳
۱۶-۵ شرایط یکسان بودن OLS و GLS	۶۱۵
۱۶-۶ تخمین‌زنندهٔ حداقل مربعات تعمیم یافتهٔ عملی (FGLS)	۶۱۷
۱۶-۷ کارایی GLS در معادلات به‌ظاهر نامرتبط	۶۱۷
۱۶-۸ آزمون قطری بودن Σ	۶۲۰
مسائل	۶۲۱
ضمیمه فصل شانزدهم: معادلات به‌ظاهر نامرتبط (SUR) در Stata	۶۲۳

فصل هفدهم: معادلات همزمان	۶۲۵
۱-۱۷ مقدمه	۶۲۵
۲-۱۷ سیستم معادلات همزمان: تعاریف و مفاهیم	۶۲۵
۳-۱۷ تخمین معادلات فرم ساختاری از روش OLS و تورش معادلات همزمان	۶۳۲
۴-۱۷ تخمین معادلات فرم خلاصه‌شده با OLS و مشکل برآورد ضرایب ساختاری	۶۳۴
۵-۱۷ شناسایی (تشخیص) معادلات فرم ساختاری	۶۳۵
۱-۵-۱۷ انواع معادلات ساختاری بر اساس قابلیت شناسایی	۶۳۵
۲-۵-۱۷ شرط درجه‌ای	۶۳۶
۳-۵-۱۷ شرط رتبه‌ای	۶۳۸
۴-۵-۱۷تشخیص معادلات با استفاده‌از محدودیت روی ماتریس‌واریانس-کوواریانس	۶۴۰
۶-۱۷ تخمین سیستم معادلات همزمان	۶۴۲
۱-۶-۱۷ روش‌های تک معادله‌ای	۶۴۲
الف) تخمین سیستم معادلات بازگشتی با روش OLS	۶۴۳
ب) تخمین معادلات همزمان با روش حداقل مربعات غیرمستقیم (ILS)	۶۴۵
ج) برآورد معادلات بیش از حد مشخص با روش 2SLS	۶۴۶
د) روش حداکثر درستنمایی با اطلاعات محدود (LIML)	۶۴۷
۲-۶-۱۷ روش‌های سیستمی	۶۴۸
الف) روش حداقل مربعات سه مرحله‌ای (3SLS)	۶۴۸
ب) برآورد سیستم معادلات همزمان با روش حداکثر درستنمایی با اطلاعات	۶۴۹
ضمیمه فصل هفدهم	۶۵۴
۱- سیستم معادلات همزمان	۶۵۴
۲- شناسایی معادلات	۶۵۶
۳- شرایط رتبه‌ای و درجه‌ای برای تشخیص معادلات	۶۵۹
۴- روش‌های تخمین سیستم معادلات همزمان	۶۶۳
۱-۴ روش‌های تک معادله‌ای	۶۶۳
الف) روش OLS	۶۶۳
ب) روش متغیرهای ابزاری (IV)	۶۶۴
ج) روش حداقل مربعات دو مرحله‌ای (2SLS)	۶۶۵
د) روش حداکثر درستنمایی با اطلاعات محدود (LIML)	۶۶۸
۲-۴ روش‌های سیستمی	۶۷۲
الف) روش 3SLS	۶۷۲
ب) روش حداکثر درستنمایی با اطلاعات کامل (FIML)	۶۷۷

مسائل ۶۸۱
ضمیمه فصل هفدهم: برآورد معادلات همزمان در Stata ۶۸۵

فصل هجدهم: مدل‌های خودرگرسیون برداری (VAR) ۶۸۷
۱-۱۸ مقدمه ۶۸۷
۲-۱۸ فرم ساختاری VAR ۶۸۷
۳-۱۸ فرم استاندارد (فرم حل‌شده) VAR ۶۸۸
۴-۱۸ VAR مقید و نامقید ۶۹۱
۵-۱۸ انتخاب طول وقفه در مدل‌های VAR ۶۹۱
۶-۱۸ شناسایی معادلات VAR ۶۹۲
۷-۱۸ تخمین حداکثر درستنمایی ۶۹۶
۸-۱۸ آزمون روابط علّی ۶۹۸
۹-۱۸ توابع واکنش ۷۰۰
۱۰-۱۸ تجزیه واریانس ۷۰۴
۱۱-۱۸ مانایی در مدل‌های VAR ۷۰۷
مسائل ۷۱۶
ضمیمه فصل هجدهم: مدل‌های VAR در Stata ۷۱۷

فصل نوزدهم: مدل‌های تصحیح خطای برداری (VECM) ۷۲۵
۱-۱۹ مقدمه ۷۲۵
۲-۱۹ نامانایی و هم‌انباشتگی در مدل‌های VAR ۷۲۵
VAR(۱) هم‌انباشته ۷۲۸
۳-۱۹هم‌انباشتگی و مدل تصحیح خطای برداری (VECM) ۷۳۰
۴-۱۹ روش جوهانسن ۷۳۱
۱-۴-۱۹ تخمین ضرایب با روش حداکثر درستنمایی ۷۳۱
۲-۴-۱۹ تعیین تعداد بردارهای هم‌انباشتگی ۷۴۵
آزمون اثر جوهانسون ۷۴۶
آزمون بزرگترین مقدار ویژه ($\lambda_{\max}$) ۷۴۷
۳-۴-۱۹ نرمال‌سازی بردارهای هم‌انباشتگی ۷۴۷
۴-۴-۱۹ شناسایی روابط هم‌انباشتگی و آزمون محدودیت‌های خطی ۷۴۸
۵-۴-۱۹ عرض از مبدأ و روند در مدل VAR و VECM ۷۵۰
مسائل ۷۶۰
ضمیمه فصل نوزدهم: مدل‌های VECM در Stata ۷۶۱

فصل بیستم: داده‌های ترکیبی ۷۶۵
۱-۲۰ مقدمه ۷۶۵
۲-۲۰ داده‌های ترکیبی ۷۶۵
۳-۲۰ مدل‌های رگرسیونی داده‌های ترکیبی ۷۷۱
۴-۲۰ علائم و عملگرها ۷۷۳
۵-۲۰ مدل تجمیعی ۷۷۸
۶-۲۰ مدل اثرات ثابت ۷۸۱
۷-۲۰ آزمون معنادار بودن اثرات ثابت ۷۸۶
۸-۲۰ تخمین‌زننده‌های درون‌گروهی و بین‌گروهی ۷۸۶
۹-۲۰ مدل اثرات تصادفی ۷۹۱
تخمین‌زنندهٔ GLS ۷۹۴
۱۰-۲۰ آزمون اثرات تصادفی ۸۰۰
۱۱-۲۰ آزمون هاسمن برای مدل اثرات تصادفی ۸۰۲
۱۲-۱۳ مدل اثرات دو طرفه ۸۰۵
۱۲-۲۰ آزمون ریشه واحد در داده‌های ترکیبی ۸۲۱
۱-۱۲-۲۰ آزمون ریشه واحد مشترک ۸۲۲
۱-۱۲-۲۰ آزمون‌های ریشه واحد مقطعی ۸۲۴
مسائل ۸۲۷
ضمیمه فصل بیستم: برآورد مدل داده‌های ترکیبی در Stata ۸۲۹

فصل بیست و یکم: متغیرهای وابسته محدود (مدل‌های لاجیت و پروبیت) ۸۳۵
۱-۲۱ مقدمه ۸۳۵
۲-۲۱ مدل‌های دو انتخابی ۸۳۶
۳-۲۱ مدل احتمال خطی (LPM) ۸۳۷
۴-۲۱ نظریه مطلوبیت تصادفی ۸۴۳
۵-۲۱ مدل پروبیت ۸۴۶
۶-۱۹ مدل لاجیت ۸۵۳
۷-۲۱ معیارهای خوبی برازش ۸۵۷
۸-۲۱ شاخص انحراف ۸۶۰
۸-۲۱ آزمون محدودیت‌ها در مدل‌های پروبیت و لاجیت ۸۶۶
مسائل ۸۶۷
ضمیمه فصل بیست و یکم: مدل‌های پروبیت و لاجیت در Stata ۸۶۸

فصل بیست و دوم: متغیرهای وابسته محدود (مدل‌های منقطع، سانسورشده و شمارشی) ۸۷۱
۱-۲۲ مقدمه ۸۷۱
۲-۲۲ توزیع‌های منقطع ۸۷۱
۳-۲۲ امید ریاضی و واریانس منقطع ۸۷۵
۴-۲۲ رگرسیون منقطع ۸۷۵
۵-۲۲ داده‌های سانسورشده ۸۸۲
۶-۲۲ توزیع نرمال سانسورشده ۸۸۲
۷-۲۲ رگرسیون سانسورشده: مدل توبیت ۸۸۴
۱-۷-۲۲ مدل توبیت ۸۸۵
۲-۷-۲۲ تفسیر نتایج مدل توبیت ۸۸۵
۸-۲۲ مدل‌های شمارشی: توزیع پوآسن ۸۹۱
۱-۸-۲۲ توزیع پوآسن ۸۹۱
۳-۸-۲۲ معیارهای خوبی برازش ۸۹۳
۴-۸-۲۲ آزمون فرضیه ۸۹۴
۵-۸-۲۲ تفسیر نتایج مدل پوآسن ۸۹۴
۶-۸-۲۲ محدودیت مدل پوآسن ۸۹۵
مسائل ۸۹۸
ضمیمه فصل بیسست و دوم: کاربردهای Stata ۸۹۹

فصل بیست و سوم: مقدمه‌ای بر اقتصاد سنجی بیزین ۹۰۳
۱-۲۳ مقدمه ۹۰۳
۲-۲۳ توزیع پیشین و پسین ۹۰۴
۳-۲۳ تخمین‌زنندهٔ بیزین ۹۱۰
۴-۲۳ تابع زیان ۹۱۱
۵-۲۳ تخمین به‌عنوان مسئلهٔ تصمیم‌گیری ۹۱۳
۶-۲۳ تخمین بیزین و کلاسیک ۹۱۵
۷-۲۳ آزمون فرضیه در روش بیزین ۹۱۹
۸-۲۳ تخمین ضرایب رگرسیون با روش بیزین ۹۲۲
۱-۸-۲۳ تابع درستنمایی ۹۲۳
۲-۸-۲۳ توزیع پسین ضرایب رگرسیون و تخمین بیزین ۹۲۷
۳-۸-۲۳ توزیع‌های پیشین با اطلاعات غیرمفید ۹۲۸
۴-۸-۲۳ توزیع‌های پیشین با اطلاعات مفید ۹۳۲

۵-۸-۲۳ تخمین نقطه‌ای و تابع زیان ۹۴۲
۶-۸-۲۳ آزمون فرضیه ۹۴۳
مسائل ۹۴۵

ضمائم ۹۴۹
ضمیمه الف: معادلات تفاضلی ۹۵۱
ضمیمه ب: مقادیر ویژه ۹۶۳
ضمیمه ج: بردارها و ماتریس‌ها ۹۶۹
ضمیمه د: جداول آماری ۹۸۵
منابع ۹۹۵
واژه‌نامه انگلیسی- فارسی ۹۹۹
واژه‌نامه فارسی- انگلیسی ۱۰۰۷
نمایه موضوعی ۱۰۱۵

فهرست کاربردهای Eviews

عنوان | صفحه

آشنایی مقدماتی با نرم‌افزار Eviews ۳۲

ایجاد فایل کاری در Eviews ۳۲

ورود داده‌ها در Eviews ۳۵

وقفه‌ها و تفاضل‌ها در Eviews ۳۶

تغییر نام متغیرها در Eviews ۳۷

متغیر زمان در Eviews ۳۷

اصلاح داده‌های وارد شده در Eviews ۳۷

مشاهده‌داده‌های وارد شده در Eviews ۳۷

نمودارها در Eviews ۳۸

تغییر دوره زمانی در Eviews ۳۹

برآورد رگرسیون ساده در Eviews ۸۰

پیش‌بینی در Eviews ۸۵

برآورد روابط معکوس در Eviews ۸۸

برآورد معادله لگاریتمی در Eviews ۸۹

برآورد تابع نمایی در Eviews	۹۰
برآورد رگرسیون انحراف از میانگین در Eviews	۹۲
برآورد رگرسیون متغیرهای استاندارد شده در Eviews	۹۳
برآورد معادله روند در Eviews	۹۵
برآورد معادله رشد در Eview	۹۶
برآورد معادلات غیرخطی درEviews	۹۶
محاسبه کوواریانس در Eviews	۹۹
محاسبه ضریب همبستگی در Eviews	۱۰۲
برآورد رگرسیون چندمتغیره در Eviews	۱۳۵
آزمون محدودیت‌ها (آزمون والد) در Eviews	۱۴۶
برآورد واریانس وایت در Eviews	۱۶۷
آزمون‌های واریانس ناهمسانی در Eviews	۱۷۳
آزمون وایت در Eviews	۱۷۷
روش GLS در Eviews	۱۸۳
آزمون دوربین- واتسون در Eviews	۱۹۴
آزمون بروش- گادفری در Eviews	۱۹۵
خودهمبستگی و مدل‌های پویا در Eviews	۲۰۰
رفع خودهمبستگی در استفاده از AR و MA در Eviews	۲۰۲
آزمون نرمال بودن در استفاده از Eviews	۲۰۶
متغیرهای مجازی در Eviews	۲۳۱
آزمون معنی‌دار بودن رگرسیون با استفاده از نسبت درستنمایی در Eviews	۲۴۷
محاسبه ضرایب همبستگی ساده در Eviews	۲۷۲
آزمون RESET رمزی در Eviews	۲۷۸
آزمون متغیرهای حذف شده در استفاده از Eviews	۲۸۴
آزمون متغیرهای نامربوط در استفاده از Eviews	۲۸۶
آزمون نقطه شکست چاو در Eviews	۲۸۹
آزمون پیش‌بینی چاو در Eviews	۲۹۲

برآوردهای بازگشتی در Eviews ۲۹۳
آزمون پیش‌بینی یک‌قدمی در Eviews ۲۹۷
آزمون مجموع تجمعی خطاها (CUSUM) در Eviews ۲۹۹
آزمون مجموع مجذور تجمعی خطاهای بازگشتی (CUSUMQ) در Eviews ۳۰۰
آزمون ثبات ضرایب با استفاده از متغیرهای مجازی در Eviews ۳۰۲
آزمون علّیت گرانجر در Eviwes ۳۰۶
متغیرهای ابزاری در Eviews ۳۳۱
برآورد مدل‌های غیرخطی در Eviews ۳۶۱
آزمون محدودیت‌ها در Eviews ۳۶۴
تخمین مدل با وقفه توزیعی در Eviews ۳۷۶
برآورد مدل وقفه خطی در Eviews ۳۷۸
برآورد مدل V معکوس در Eviews ۳۸۰
روش آلمون در Eviews ۳۸۳
تبدیل کویک در Eviews ۳۹۲
الگوی وقفهٔ پاسکال در Eviews ۳۹۴
برآورد مدل ECM در Eviews ۴۰۴
ایجاد فرایند تصادفی محض در Eviews ۴۱۰
برآورد ضرایب خودهمبستگی در Eviwes ۴۱۲
ایجاد فرایند MA در Eviews ۴۲۰
ایجاد فرایند AR در Eviews ۴۳۰
برآورد مدل‌های ARMA در Eviews ۴۴۱
بازبینی مدل با استفاده از Eviews ۴۴۳
معیارهای اطلاعات در Eviwes ۴۴۵
پیش‌بینی سری‌های زمانی در Eviews ۴۵۳
آزمون ریشه واحد در Eviews ۴۷۴
آزمون فیلیپس- پرون در Eviews ۴۷۹
آزمون ریشه واحد در حالت شکست ساختاری در Eviews ۴۸۳

آزمون هم‌انباشتگی در Eviews ۴۹۰
آزمون ریشه واحد فصلی در Eviews ۵۲۸
آزمون ARCH در Eviews ۵۴۳
تخمین مدل GARCH در Eviews ۵۴۹
برآورد مدل GJR در Eviews ۵۵۳
برآورد مدل TARCH در Eviews ۵۵۵
برآورد مدل EGARCH در Eviews ۵۵۷
برآورد GARCH-M در Eviews ۵۶۳
برآورد GARCH با اجزاء موقتی و دائمی در Eviews ۵۶۴
پیش‌بینی واریانس شرطی در Eviews ۵۶۷
تخمین GARCH چند متغیره در Eviews ۵۷۳
آزمون تغییر جهت در Eviews ۵۹۲
برآورد مدل SUR در Eviews ۶۲۰
تخمین معادلات همزمان در Eviews ۶۴۹
برآورد مدل VAR در Eviews ۷۱۲
برآورد مدل VECM در Eviews ۷۵۳
تخمین مدل داده‌های ترکیبی در Eviews ۸۰۸
آزمون ریشه واحد داده‌های ترکیبی در Eviews ۸۲۴
برآورد مدل LPM در Eviews ۸۴۲
تخمین مدل پروبیت و لاجیت در Eviews ۸۶۴
برآورد رگرسیون منقطع در Eviews ۸۷۸
برآورد رگرسیون سانسور شده (مدل توبیت) در Eviews ۸۹۰
برآورد مدل پوآسن در Eviews ۸۹۶

فهرست کاربردهایStata

عنوان	صفحه
آشنایی مقدماتی با نرم‌افزار Stata	۴۲
ایجاد فایل کاری و ورود داده‌ها در Stata	۴۲
مشاهده و اصلاح داده های وارد شده در Stata	۴۳
توصیف کلی متغیرها در Stata	۴۳
معرفی منوی Statistics در Stata	۴۵
ایجاد متغیر جدید در Ststs	۴۶
متغیر زمان در Stata	۴۷
نمودارها در Stata	۴۷
سایر فرمان‌ها در Stata	۴۹
افزودن دامنه در Stata	۴۹
سایر علائم در Stata	۴۹
لحاظ نکردن برخی از مشاهدات در Ststa	۵۰
حذف یک متغیر در Stata	۵۰
نشان‌گذاری متغیرها در Stata	۵۰

برآورد رگرسیون ساده در Stata ۱۱۰
محاسبهٔ مقادیر تخمینی و باقیمانده‌ها در Stata ۱۱۲
پیش‌بینی در Stata ۱۱۲
محاسبهٔ ماتریس واریانس- کوواریانس ضرایب در Stata ۱۱۴
برآورد معادلات غیرخطی در Stata ۱۱۵
محاسبهٔ ضرایب همبستگی درStata ۱۱۶
برآورد معادلات رگرسیون چند متغیره در Stata ۱۵۵
آزمون محدودیت‌ها (آزمون والد) در Stata ۱۵۶
تخمین رگرسیون مقید در Stata ۱۵۷
آزمون‌های واریانس ناهمسانی در Stata ۲۱۲
برآورد واریانس مستحکم وایت در Stata ۲۱۴
روش GLS در Stata ۲۱۴
آزمون خودهمبستگی در Ststs ۲۱۵
آزمون دوربین- واتسون در Stata ۲۱۵
آزمون نرمال بودن در Stata ۲۱۵
متغیرهای مجازی در Stata ۲۳۴
آزمون حذف متغیرهای مهم در Stata ۳۱۰
آزمون متغیرهای زائد در Stata ۳۱۱
آزمون چاو در Stata ۳۱۱
آزمون علیت گرانجر در Stata ۳۱۲
متغیرهای ابزاری در Stata ۳۳۴
آزمون هاسمن در Stata ۳۳۵
برآورد معادلات غیرخطی در Stata ۳۶۹
تخمین مدل‌های با وقفه توزیعی در Stata ۴۰۶
محاسبهٔ ضرایب خودهمبستگی در Stata ۴۵۶
برآورد مدل‌های ARIMA در Stata ۴۵۸
پیش‌بینی مدل‌های ARIMA در Stata ۴۶۰

آزمون ریشه واحد ر Stata	۴۶۸
آزمون هم‌انباشتگی در Stata	۵۰۰
آزمون ریشه واحد فصلی در Stata	۵۳۱
آزمون ARCH در Stata	۵۸۰
تخمین مدل‌های GARCH در Stata	۵۸۱
برآورد مدل GARCH(p,q) در Stata	۵۸۲
برآورد مدل TARCH در Stata	۵۸۴
برآورد مدل EGARCH در Stata	۵۸۵
برآورد GARCH-M در Stata	۵۸۵
برآورد مدل SUR در Stata	۶۲۳
تخمین معادلات همزمان در Stata	۶۸۵
برآورد مدل VAR در Stata	۷۱۷
برآورد مدل VECM در Stata	۷۶۱
تخمین مدل داده‌های ترکیبی در Stata	۸۲۹
آزمون ریشه واحد داده‌های ترکیبی در Stata	۸۳۴
تخمین مدل پروبیت و لاجیت در Stata	۸۶۸
برآورد رگرسیون منقطع در Stata	۸۹۹
برآورد رگرسیون سانسور شده (مدل توبیت) در Stata	۹۰۰
برآورد مدل پوآسن در Stata	۹۰۱

مقدمه

کتاب حاضر در راستای تدوین مجموعه‌ای نسبتاً جامع از مباحث اقتصاد سنجی برای دانشجویان کارشناسی و کارشناسی ارشد رشته‌های اقتصاد و مالی می‌باشد. نسخه‌های قبلی این کتاب صرفاً شامل بخشی از مباحث اقتصاد سنجی بود که به‌صورت مقدماتی و آزمایشی به چاپ رسید. اکنون مجموعه کامل‌تری از این مباحث تهیه شده است که می‌تواند نیازهای مخاطبان بیشتری را تأمین نماید. مباحث کتاب به دو بخش مقدماتی و پیشرفته تقسیم شده است که به‌ترتیب شامل جلد ۱ و ۲ می‌باشد. در تدوین این مجموعه سعی شده که اولاً مبانی نظری به‌طور کامل ارائه گردد، ثانیاً ویژگی کاربردی مباحث حفظ شود و ثالثاً همراه با کاربرد نرم‌افزارها باشد که در این خصوص از دو نرم افزار رایج، Eviews و Stata استفاده شده است.

مباحث اقتصاد سنجی دامنه وسیعی از مطالب را از سطوح مقدماتی تا پیشرفته و از مباحث کاربردی تا نظری شامل می‌شود. با توجه به گستردگی مباحث، تنظیم کتابی که بتواند نیاز مخاطبین آن را تأمین نماید، بسیار دشوار است. با توجه به این نکات، در این کتاب سعی بر آن است که با رعایت حد میانه‌ای از تئوری و کاربرد، نیاز مخاطبین تأمین گردد. در این کتاب، تا حدود زیادی سعی شده تا مباحث اولیه با فشردگی بیشتری بیان شود تا امکان ارائه مباحث جدیدتر فراهم گردد. از طرف دیگر، چون برآورد معادلات و تحلیل مباحث، بدون استفاده از نرم‌افزار امکان‌پذیر نیست، لذا در پایان هر مبحث، کاربرد آن با استفاده از نرم‌افزار Eviews و سپس در پایان هر فصل، کاربردهای نرم‌افزار Stata نیز ارائه شده است. علاوه بر این، داده‌های مورد استفاده

همراه با نرم‌افزار Eveiws و Stata نیز در قالب یک CD ارائه شده است تا امکان استفادهٔ همزمان از کتاب و نرم‌افزار فراهم گردد. همچنین در هر جایی که کاربردهای Eviews و Stata آمده است، نام فایل مورد استفاده نیز درج شده تا امکان انجام مستقیم آنها فراهم شود.

در تنظیم این کتاب از منابع مختلفی استفاده شده است که البته در مواردی، کتاب اقتصاد سنجی مالی بروکس (۲۰۰۱ و ۲۰۰۸) و اقتصاد سنجی گرین (۲۰۱۲) بیش از بقیه، مورد استفاده قرار گرفته است.

این کتاب در دو جلد و ۲۳ فصل تنظیم شده است. فصل اول به کلیاتی راجع‌به توزیع‌های آماری و آشنایی مقدماتی با نرم‌افزارEviews و Stata اختصاص دارد. در این فصل مفاهیم و توزیع‌های مهم آماری که همواره در اقتصادسنجی مورد استفاده قرار می‌گیرد، ارائه شده است. همچنین بخشی نیز به نکات اولیه و مقدماتی برای آشنایی با نرم‌افزار Eviews و Stata اختصاص یافته است. فصل دوم اختصاص به مباحث پایه‌ای رگرسیون ساده دارد که در همه کتاب‌های اقتصادسنجی به چشم می‌خورد. در اینجا سعی شده تا مباحث رگرسیون را از توزیع‌های دو متغیره شروع کنیم. این شیوه کمک می‌کند تا مفهوم رگرسیون به‌عنوان امید ریاضی شرطی به‌خوبی توصیف شود. در فصل سوم نیز مباحث مذکور برای حالت‌های چند متغیره ارائه شده است. فصل چهارم اختصاص به نقض فروض کلاسیک و آزمون‌های مربوطه دارد.

در فصل پنجم، متغیرهای مجازی و نحوه استفاده از آنها در مباحث مختلف، ارائه شده است.

در فصل ششم سایر موضوعاتی که در کارهای تجربی به‌کار می‌روند، ارائه شده است. این موضوعات عمدتاً شامل تصریح مدل و آزمون‌های مربوطه می‌باشد.

فصل هفتم اختصاص به روش حداکثر درستنمایی و موضوعات و آرمون‌های پیرامون آن دارد. به‌ویژه سه آزمون معروف نسبت درستنمایی، والد و ضریب لاگرانژ به‌تفصیل بحث شده‌اند.

فصل هشتم اختصاص به متغیرهای ابزاری دارد که شامل مباحثی در خصوص روش IV و روش 2SLS می‌باشد.

فصل نهم به رگرسیون غیر خطی و روش‌های تخمین و الگوریتم‌های مربوطه دارد که به‌تفصیل بحث شده‌اند.

در فصل دهم الگوهای پویای تأخیری بحث شده است. در این فصل، روش‌های مختلف تخمین الگوهای پویا که با وقفه‌های توزیعی همراه هستند ارائه شده است. در پایان این فصل الگوهای ARDL نیز مورد بحث قرار گرفته است.

فصل‌های یازدهم و دوازدهم و سیزدهم به سری‌های زمانی یک متغیره اختصاص دارد. در فصل یازدهم مدل‌های ARIMA و در فصل دوازدهم مسائل مربوط به نامانایی، ریشه واحد و هم‌انباشتگی بحث شده است. فصل سیزدهم نیز اختصاص به سری‌های زمانی فصلی دارد که خصوصیات آنها با سری‌های زمانی معمولی متفاوت است.

در فصل چهاردهم مدل‌های تغییرپذیری ارائه شده است. این مدل‌ها شامل انواع مدل‌های ARCH و GARCH می‌باشند. در فصل پانزدهم نیز نوع خاصی از تغییرات سری‌های زمانی بررسی شده است که موسوم به مدل‌های تغییر جهت هستند.

فصل‌های شانزده تا نوزده اختصاص به سیستم معادلات دارد. در فصل شانزدهم معادلات به‌ظاهر نامرتبط و در فصل هفدهم سیستم معادلات همزمان و انواع روش‌های تخمین آنها ارائه شده است. در فصل هجدهم و نوزدهم نیز سری‌های زمانی چند متغیر تحت عنوان مدل‌های خود رگرسیون برداری (VAR) مدل‌های تصحیح خطای برداری (VECM) ارائه شده‌اند.

داده‌های ترکیبی یا مدل‌های پانل در فصل بیستم ارائه شده است. در این فصل مدل‌های تجمیعی، اثرات ثابت و اثرات تصادفی و آزمون‌ها و روش‌های مربوطه به‌تفصیل بحث شده است.

فصل‌های بیست و یکم و بیست و دوم در مورد متغیرهای محدود بحث می‌کند. در این دو فصل مباحثی در خصوص متغیرهای وابستهٔ محدود از قبیل مدل‌های لاجیت و پروبیت، مدل پوآسن، مدل‌های سانسور شده و منقطع بحث شده است.

فصل بیست و سوم مباحث مقدماتی در مورد اقتصاد سنجی بیزین رائه می‌کند. در این فصل ابتدا مباحث اولیه مانند توزیع‌های پسین و پیشین، تخمین‌زنندهٔ کلاسیک و بیزین و روش تخمین بیزین در معادلات رگرسیون ارائه شده است.

به‌هرحال گستردگی مباحث اقتصاد سنجی، امکان ارائه همهٔ مطالب را میسر نمی‌سازد. بدیهی است که هر چند سعی شده تا مطالب بیشتری ارائه شود ولی هنوز کاستی‌های زیادی وجود دارد. حجم مطالب و تنوع آن به‌گونه‌ای است که امکان وجود اشتباهات را زیاد می‌کند و بدیهی است که کاستی‌ها و نواقص این کتاب زیاد خواهد بود که بابت آن از اساتید محترم و دانشجویان گرامی عذر تقصیر می‌خواهم. بی‌تردید ارائه نظرات شما می‌تواند در بهبود و ارتقای مطالب کتاب، راهگشا باشد. منتظر دریافت نظرات ارزشمند و اصلاحی شما از طریق ایمیل souri.econometrics@yahoo.com هستم. در پایان لازم می‌دانم از همه همکارن و دانشجویانی که نظرات اصلاحی خود را به اینجانب منعکس نموده‌اند تشکر کنم. همچنین لازم است که از صبر و مساعدت همسرم قدردانی کنم که با تقبل همهٔ مسئولیت‌های زندگی، مرا در نگارش این مجموعه یاری نمودند و همچنین ویرایش و بازخوانی مطالب و صفحه‌آرایی آن را به‌عهده گرفتند، لذا این اثر را به ایشان تقدیم می‌کنم.

علی سوری

بهمن ۱۳۹۲

فصل اول

مروری بر مفاهیم پایه‌ای آمار و نرم‌افزارهای Eviews و Stata

۱-۱ مقدمه

در این فصل ابتدا مروری به برخی از مفاهیم پایه‌ای و توابع آماری که استفاده گسترده‌ای در اقتصادسنجی دارند، خواهیم داشت. سپس به معرفی اجمالی نرم‌افزارهای Eviews و Stata و اصول مقدماتی آنها می‌پردازیم. سایر کاربردهای این نرم‌افزار، در فصول بعدی و همراه با مباحث مربوطه، ارائه خواهند شد.

۲-۱ نمونه تصادفی

نمونه تصادفی قسمتی از «جامعه آماری» است که به‌طور تصادفی انتخاب می‌شود. در اینجا لازم است دو مفهوم را از هم متمایز کنیم که یکی «نمونه مشاهده شده» و دیگری «نمونه تصادفی» است. «نمونه مشاهده شده» فقط شامل اعداد و ارقامی است که بعد از نمونه‌گیری به‌دست می‌آیند. اما «نمونه تصادفی» به مرحلهٔ قبل از نمونه‌گیری مربوط می‌شود و نمی‌دانیم که چه مقادیری را مشاهده خواهیم کرد.

مثال ۱-۱: فرض کنید که جامعه آماری شامل ۳ دانشجو است که نمرات آنها در درس آمار شامل ۱۰، ۱۲ و ۱۵ می‌باشد. لذا $N=۳$ بوده و می‌خواهیم نمونه‌ای به حجم $n=۲$ با

روش جایگزینی انتخاب کنیم. بدیهی است که یکبار نمونه‌گیری را انجام خواهیم داد و بعد از انجام نمونه‌گیری فقط دو عدد (نمره) به‌دست خواهد آمد. اما قبل از انجام نمونه‌گیری نمی‌دانیم که کدام دو عدد (نمره) را به‌دست خواهیم آورد. لذا نمونهٔ تصادفی شامل متغیرهای تصادفی $(X_۱, X_۲)$ و نمونهٔ مشاهده شده شامل اعداد یا مقادیر $(x_۱, x_۲)$ خواهد بود. توجه شود که $X_۱$ بیانگر انتخاب اول و $X_۲$ بیانگر انتخاب دوم می‌باشد. تمامی نتایج ممکن برای $(X_۱, X_۲)$ به‌صورت زیر می‌باشد:

(۱۰,۱۰) (۱۲,۱۰) (۱۵,۱۰)
(۱۰,۱۲) (۱۲,۱۲) (۱۵,۱۲)
(۱۰,۱۵) (۱۵,۱۲) (۱۵,۱۵)

بنابراین ۹ (=$۳^۲$) نتیجهٔ ممکن خواهیم داشت. بدیهی است که فقط یک‌بار نمونه‌گیری انجام می‌شود و لذا «نمونهٔ مشاهده شده» فقط شامل یکی از نمونه‌های فوق خواهد بود.

بنابراین، نمونه تصادفی را به‌صورت زیر تعریف می‌کنیم:

«نمونه تصادفی مجموعه‌ای از n متغیر تصادفی $(X_۱, X_۲, \cdots, X_n)$ است که مستقل از هم بوده و توزیع آنها با توزیع X در جامعه آماری یکسان است.»

۳-۱ توزیع مشترک متغیرهای نمونه یا توزیع مشترک نمونه تصادفی

یکی از نکات مهم در رابطه با نمونه تصادفی این است که هر متغیر نمونه (یعنی X_i) دقیقاً همان ویژگی‌های متغیر تصادفی X در جامعه اصلی را دارد. به‌عنوان مثال فرض کنید جامعه آماری شامل ۱۰ نفر (N=۱۰) باشد که متغیر تصادفی X بیانگر بی‌سوادی اعضای این جامعه است:

$X = ۱$ بی‌سواد

$X = ۰$ باسواد

بنابراین اگر نسبت بی‌سوادی برابر با p باشد، در این صورت اگر یک نفر را از این جامعه انتخاب کنیم، احتمال اینکه بی‌سواد باشد (یا احتمال اینکه x=۱ باشد) برابر با p است و در غیر این صورت برابر با q یا p-۱ است.

x	۰	۱
$P(x)$	q	p

و یا می‌توان آن را به‌صورت تابع احتمال دونقطه‌ای نشان داد:

$$P(x)=p^{x}q^{۱-x} \quad ; \quad x=۰,۱$$

حال اگر از این جامعه، نمونه‌ای به حجم n=۲ انتخاب کنیم، در این صورت نمونه تصادفی شامل $(X_۱,X_۲)$ است. حال سؤال این است که توزیع $X_۱$ چگونه می‌باشد؟

اولاً مقادیر $X_۱$ در نمونه، دقیقاً همان مقادیری است که X در جامعه اصلی اختیار می‌کند. زیرا مقدار مشاهدهٔ اول یعنی $X_۱$ یا ۰ است و یا ۱. اما احتمال اینکه $X_۱=۰$ باشد برابر است با احتمال اینکه $X=۰$ باشد که هر دو بستگی به این دارند که چه نسبتی از جامعه بی‌سواد نمی‌باشد (یعنی q). بنابراین، توزیع $X_۱$ دقیقاً همان توزیع X است. به‌عبارت دقیق‌تر متغیر تصادفی $X_۱$ با متغیر تصادفی X هیچ‌گونه تفاوتی ندارد.

$$P(x_۱)=p^{x_۱}q^{۱-x_۱} \quad ; \quad x_۱=۰,۱$$

یا

$x_۱$	۰	۱
$P(x_۱)$	q	p

همین استدلال را برای $X_۲$ نیز می‌توان به‌کار برد.

بنابراین اگر در حالت کل، نمونه تصادفی به حجم n باشد، در این صورت هر یک از متغیرهای تصادفی $X_۱$، $X_۲$، ... و X_n دقیقاً همان توزیع X در جامعه اصلی را خواهند داشت. در مثال فوق، هر یک از X_iها توزیع دونقطه‌ای خواهند داشت:

$$P(x_i)=p^{x_i}q^{۱-x_i} \quad x_i=۰,۱ \quad , \quad i=۱,۲,\cdots,n$$

مثال ۲-۱: فرض کنید که در جامعه آماری متغیر تصادفی X بیانگر قد دانشجویان یک دانشگاه است که دارای توزیع نرمال با میانگین ۱۶۹ و واریانس ۶۴ می‌باشد:

$$X \sim (\mu=۱۶۹,\sigma^{۲}=۶۴)$$

اگر از این جامعه نمونه‌ای تصادفی به حجم n انتخاب کنیم هر یک از متغیرهای تصادفی $X_۱$، $X_۲$، ... و X_n دارای توزیع نرمال خواهند بود:

$$E(X_i)=E(X)=\mu \quad , \quad \text{var}(X_i)=\text{var}(X)=\sigma^{۲}$$

$$X_i \sim (\mu=۱۶۹,\sigma^{۲}=۶۴)$$

در حالت کلی، برداری از متغیرهای تصادفی یعنی $X_۱$، $X_۲$،... و X_n را داریم که مقادیر آنها با $x_۱$، $x_۲$، ... و x_n نشان داده می‌شود. حال توزیع مشترک یا احتمال مشترک این نمونه تصادفی برابر است با:

$$P(X_۱ = x_۱, X_۲ = x_۲, \cdots, X_n = x_n) \tag{۱-۱}$$

که بیانگر این است که چقدر احتمال دارد در نمونه‌گیری تصادفی، مقادیر $x_۱$، $x_۲$، ... و x_n مشاهده گردد. اما از آنجا که این متغیرها، مستقل از هم می‌باشند لذا احتمال مذکور برابر است با:

$$P(X_۱ = x_۱, X_۲ = x_۲, \cdots, X_n = x_n) = P(X_۱ = x_۱)P(X_۲ = x_۲)\cdots P(X_n = x_n) = \prod_{i=۱}^{n} P(X_i = x_i) = \prod_{i=۱}^{n} P(x_i) \tag{۱-۲}$$

در حالت پیوسته به‌جای تابع احتمال، از تابع چگالی استفاده می‌شود که در آن صورت تابع چگالی مشترک به‌دست می‌آید.

مثال ۳-۱: در مثال ۱-۱ اگر حجم نمونه n باشد، در این صورت توزیع مشترک متغیرهای نمونه عبارت است از:

$$P(X_۱ = x_۱, \cdots, X_n = x_n) = P(x_۱)\cdots P(x_n) = p^{\sum_{i=۱}^{n} x_i} q^{n-\sum_{i=۱}^{n} x_i}$$

مثال ٤-۱: در مثال ۲-۱ توزیع مشترک متغیرهای نمونه عبارت است از:

$$P(X_۱ = x_۱, \cdots, X_n = x_n) = f_{X_۱}(x_۱)\cdots f_{X_n}(x_n)$$

$$= f_X(x_۱)\cdots f_X(x_n) = \frac{۱}{(۲\pi)^{\frac{n}{۲}} \sigma^n} e^{-\frac{\sum_i (x_i - ۱۶۹)^۲}{۱۲۸}}$$

٤-۱ تابعی از متغیرهای نمونه (آماره)

از آنجا که $X_۱$، $X_۲$، ... و X_n متغیرهای تصادفی هستند، لذا هر تابعی از این متغیرها نیز یک متغیر تصادفی است. همان‌طور که خواهیم دید، این توابع نقش مهمی در مباحثی چون تخمین و آزمون فرضیه دارند. شکل کلی چنین توابعی عبارت است از:

$$f(X_۱, X_۲, \cdots, X_n)$$

توجه شود که این تابع، فقط تابعی از متغیرهای نمونه است شامل هیچیک از پارامترهای توزیع X نیست. برخی از انواع مهم این توابع عبارتند از:

$$Y = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\widetilde{S}^2 == \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

$$S^2 == \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

دو تابع نمونه‌ای یا آمارهٔ $\overline{X}$ و S^2 بیش از بقیه، مورد استفاده می‌باشد که اولی میانگین نمونه و دیگری واریانس نمونه است.

بر اساس هریک از توابع فوق می‌توان توابع دیگری را نیز تعریف کرد که در ادامه به آنها خواهیم پرداخت. از آنجا که هر یک از این توابع، متغیرهای تصادفی هستند لذا برای آنها می‌توان تابع احتمال یا چگالی را نیز به‌دست آورد.

۵-۱ تخمین

موضوع تخمین راجع به برآورد پارامترهای جامعهٔ اصلی است. در اینجا فرض می‌کنیم که توزیع X در جامعه اصلی معلوم است ولی پارامترهای آن نامعلوم هستند. به‌عنوان مثال فرض کنید که توزیع X در جامعه اصلی، نرمال است اما پارامترهای آن یعنی μ و σ^2 مجهول می‌باشند.

۱-۵-۱ تخمین و تخمین‌زننده

فرض کنید نسبت بی‌سوادی در جامعه‌ای دارای توزیع دونقطه‌ای است که به‌صورت زیر نشان داده می‌شود:

$$P(x) = p^x q^{1-x} \quad ; \quad x = 0,1$$

این توزیع یک پارامتر دارد که p (یعنی نسبت بی‌سوادی) است. بنابراین توزیع X معلوم است اما پارامتر این توزیع یعنی p، نامعلوم می‌باشد. حال با استفاده از اطلاعات نمونه می‌خواهیم p را برآورد نماییم. فرض کنید نمونه‌ای به حجم n انتخاب می‌کنیم. بعد از انجام نمونه‌گیری، مشاهدات ما شامل مقادیر x_1، x_2، ... و x_n است که بر اساس آن، تخمین p عبارت است از:

$$\frac{x_1+x_2+...+x_n}{n}$$

هر کدام از x_iها فقط ۰ یا ۱ هستند. اگر همگی صفر باشند در این صورت تخمین p برابر با ۰ و اگر همگی برابر با ۱ باشند تخمین p برابر با ۱ خواهد بود. حال اگر برخی از x_iها برابر با ۰ و برخی نیز برابر با ۱ باشند، در این صورت تخمین p بین ۰ و ۱ خواهد بود. در حالت کلی اگر آن تعداد از x_iها را که برابر با ۱ هستند با m نشان دهیم، در این صورت $m=x_1+x_2+...+x_n$ بوده و لذا خواهیم داشت:

$$\frac{m}{n}=\frac{x_1+x_2+...+x_n}{n}$$

بنابراین تخمین p معادل با $\frac{m}{n}$ است. به همین ترتیب اگر بحث خود را معطوف به بعد از نمونه‌گیری کنیم، در این صورت مقدار m معلوم بوده و بر اساس آن، تخمین p به‌دست خواهد آمد. اما قبل از نمونه‌گیری، m یک متغیر تصادفی خواهد بود، زیرا مقدار آن بعد از انجام آزمایش (نمونه‌گیری) معلوم خواهد شد و لذا m یا $\frac{m}{n}$ تابعی از متغیرهای نمونه است:

$$\hat{p}=\frac{m}{n}=\frac{X_1+X_2+...+X_n}{n}$$

به چنین تابعی «تخمین‌زننده» و یا «برآوردکننده» یا «برآوردگر» و مقدار حاصل از این تابع را که بعد از نمونه‌گیری به‌دست می‌آید «تخمین» یا «برآورد» می‌گوییم. بنابراین تخمین‌زنندهٔ هر پارامتری مانند θ را با $\hat{\theta}$ نشان می‌دهیم که تابعی از متغیرهای نمونه است:

$$\hat{\theta}=f(X_1,X_2,\cdots,X_n) \tag{۱-۳}$$

اما بدیهی است که برای پارامتر θ نمی‌توان هر تابع دلخواهی را به‌عنوان تخمین‌زننده تعریف نمود، بلکه هر تخمین‌زننده باید دارای خواصی باشد که در ادامه به آنها می‌پردازیم. برای به‌دست

آوردن تخمین‌زننده‌هایی که دارای خواص مناسبی باشند، روش‌هایی وجود دارد که ابتدا این روش‌ها را بررسی می‌کنیم.

۲-۵-۱ روش‌های تخمین

برای تخمین پارامترهای مورد نظر، روش‌هایی وجود دارد که هر یک دارای ویژگی‌ها و خواص خود هستند. در اینجا به بررسی این روش‌ها می‌پردازیم.

روش گشتاورها

روش گشتاورها بر مبنای این استدلال ساده قرار دارد که گشتاورهای هر توزیع، تابعی از پارامترهای آن توزیع می‌باشد و لذا می‌توان با حل چنین معادلاتی به تخمین پارامترها بر حسب گشتاورها رسید. به‌عنوان مثال اگر توزیعی فقط پارامتر θ را داشته باشد، در این صورت اولین گشتاور که همان میانگین می‌باشد، عبارت است از:

$$m_1 = E(X) = g(\theta)$$

حال اگر به‌جای m_1 مقدار آن را بر حسب اطلاعات نمونه قرار دهیم، خواهیم داشت:

$$\hat{m}_1 = g(\hat{\theta})$$

که $\hat{m}_1$ گشتاور مرتبه اول نمونه و $\hat{\theta}$ تخمین θ می‌باشد. از حل این معادله خواهیم داشت:

$$\hat{\theta} = h(\hat{m}_1)$$

حال اگر توزیعی دارای دو پارامتر باشد، در این صورت گشتاورهای مرتبه اول و دوم را محاسبه می‌کنیم:

$$m_1 = E(X) = g_1(\theta_1, \theta_2)$$

$$m_2 = E(X^2) = g_2(\theta_1, \theta_2)$$

و با جایگذاری از اطلاعات نمونه به‌جای گشتاورهای جامعه، خواهیم داشت:

$$\hat{m}_1 = g_1(\hat{\theta}_1, \hat{\theta}_2)$$

$$\hat{m}_2 = g_2(\hat{\theta}_1, \hat{\theta}_2)$$

با حل این دو معادله برای پارامترهای $\hat{\theta}_1$ و $\hat{\theta}_2$ خواهیم داشت:

$$\hat{\theta}_1 = h_1(\hat{m}_1, \hat{m}_2)$$

$$\hat{\theta}_2 = h_2(\hat{m}_1, \hat{m}_2)$$

مثال ۵-۱: فرض کنید X دارای توزیع یکنواخت است:

$$f(x)=\frac{۱}{b} \qquad ۰<x<b$$

پارامتر b را با استفاده از روش گشتاورها به‌دست آورید.

$$m_۱=E(X)=\int_۰^b x\frac{۱}{b}dx=\frac{x^۲}{۲b}\Big|_۰^b=\frac{b}{۲}$$

بنابراین به‌جای $m_۱$ از $\hat{m}_۱$ و به‌جای b از $\hat{b}$ استفاده می‌کنیم.

$$\hat{m}_۱=\frac{\hat{b}}{۲} \Rightarrow \hat{b}=۲\hat{m}_۱$$

که $\hat{b}$ تابعی از $\hat{m}_۱$ می‌باشد.

مثال ۶-۱: فرض کنید که X دارای یکنواخت باشد:

$$f(x)=\frac{۱}{b-a}\ ;\ a<x<b$$

پارامترهای a و b را با استفاده از روش گشتاورها تخمین بزنید.

$$m_۱=\int_a^b x\frac{۱}{b-a}dx=\frac{a+b}{۲}$$

$$m_۲=\int x^۲\frac{۱}{b-a}dx=\frac{b^۲+ab+a^۲}{۳}$$

که می‌توان آنها را به‌صورت زیر نوشت:

$$\hat{m}_۱=\frac{\hat{a}+\hat{b}}{۲}$$

$$\hat{m}_۲=\frac{\hat{b}^۲+\hat{a}\hat{b}+\hat{a}^۲}{۳}$$

با حل این معادلات، نتیجه می‌شود:

$$\hat{a}=\hat{m}_۱+\sqrt{۳(\hat{m}_۲-\hat{m}_۱^۲)}$$

$$\hat{b}=\hat{m}_۱-\sqrt{۳(\hat{m}_۲-\hat{m}_۱^۲)}$$

روش حداکثر درستنمایی

روش حداکثر احتمال یا حداکثر درستنمایی یکی از روش‌های مناسب برای برآورد پارامترها می‌باشد. مبنای این روش بر اساس استدلال زیر است:

فرض کنید X دارای توزیع معینی باشد که فقط یک پارامتر (θ) دارد که آن را به‌صورت $P(x,\theta)$ نشان می‌دهیم. حال وقتی نمونه‌گیری را انجام دهیم، نتایج حاصل از نمونه‌گیری شامل مقادیر $x_۱$، $x_۲$، ... و x_n می‌باشد. احتمال مشاهده شدن این نمونه، عبارت است از:

$$\begin{aligned} P(X_۱ = x_۱,...,X_n = x_n) &= P(X_۱ = x_۱)...P(X_n = x_n) \\ &= P(x_۱,\theta)...P(x_n,\theta) \end{aligned} \qquad (۱-۴)$$

بدیهی است که مقدار این احتمال بستگی به مقدار θ دارد، زیرا مقادیر متغیرهای تصادفی $X_۱$، $X_۲$، ... و X_n معلوم شده‌اند و به‌جای آن، مقادیر $x_۱$، $x_۲$، ... و x_n را قرار داده‌ایم. مقدار پارامتر θ مجهول است که تعیین‌کننده مقدار این احتمال می‌باشد. لذا مقدار این احتمال در رابطهٔ (۴-۱) تابعی از θ است که آن را با $L(\theta)$ نشان می‌دهیم. $L(\theta)$ بیانگر «احتمال نمایان شدن نمونه مورد نظر» می‌باشد که به آن «تابع درستنمایی» می‌گوییم.

حال سؤال این است که از بین تمامی نمونه‌هایی که می‌توانستند نمایان شوند، چرا فقط این مقادیر (یعنی $x_۱$، $x_۲$، ... و x_n) مشاهده شده‌اند. دلیل آن ساده است، زیرا این مشاهدات، بیشترین احتمال یا شانس را برای نمایان شدن داشته‌اند. پس باید $L(\theta)$ بیانگر حداکثر احتمال باشد. لذا مقدار θ که بر حسب مشاهدات نمونه تعیین می‌شود باید به‌گونه‌ای باشد که احتمال مذکور را حداکثر نماید. برای حداکثر شدن $L(\theta)$، شرط زیر را به‌کار می‌بریم:

$$\frac{\partial L}{\partial \theta} = ۰ \quad \Rightarrow \quad \hat{\theta} = g(x_۱, x_۲,..,x_n) \qquad (۱-۵)$$

معمولاً بهتر است که ابتدا از $L(\theta)$ لگاریتم گرفته و سپس نسبت به θ مشتق بگیریم، زیرا آن مقدار از θ که لگاریتم L را حداکثر نماید، L را نیز حداکثر خواهد نمود.[۱]

۱- اگر $f(x)$ دارای نقطه حداکثر باشد شرط لازم عبارت است از:

$$f'(x) = ۰$$

اما اگر $\ln f(x)$ را حداکثر کنیم، خواهیم داشت:

$$\frac{\partial \ln f(x)}{\partial x} = ۰ \Rightarrow \frac{f'(x)}{f(x)} = ۰ \Rightarrow f'(x) = ۰$$

بنابراین در هر دو حالت، شرط لازم به‌صورت $f'(x)=۰$ می‌باشد.

مثال ۷-۱: فرض کنید X توزیع دونقطه‌ای دارد:

$$P(x) = p^{x}(1-p)^{1-x}$$

از این جامعه، نمونه‌ای به حجم n انتخاب می‌شود. تخمین پارامتر p با استفاده از روش حداکثر درستنمایی عبارت است از:

$$\begin{aligned}
L(p) &= P(x_1,p)\ldots P(x_n,p) = p^{x_1}(1-p)^{1-x_1}\ldots p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum x_i}(1-p)^{n-\sum x_i} \\
\ln L(p) &= \sum x_i \ln p + \left(n-\sum x_i\right)\ln(1-p) \\
\frac{\partial \ln L}{\partial p} &= \sum x_i \frac{1}{p} + \left(n-\sum x_i\right)\frac{-1}{1-p} = 0 \Rightarrow \hat{p} = \frac{\sum x_i}{n} = \bar{x}
\end{aligned}$$

بنابراین تابع $L(p)$ به‌ازای $\hat{p} = \frac{\sum X_i}{n} = \bar{X}$ حداکثر خواهد شد.

بدیهی است که اگر تعداد پارامترها بیشتر باشد مثلاً دو پارامتر θ_1 و θ_2 داشته باشیم، در این صورت تابع درستنمایی تابعی از θ_1 و θ_2 است که نسبت به هر دو مشتق گرفته و تخمین‌زننده آنها را به‌دست می‌آوریم.

روش حداقل مربعات معمولی

این روش یکی از روش‌های مهم در تخمین پارامترها می‌باشد که عمدتاً برای تخمین معادلات رگرسیون به‌کار می‌رود. جزئیات این روش در فصل‌های بعدی ارائه خواهد شد.

۳-۵-۱ خواص تخمین‌زننده‌ها

از آنجا که برای تخمین یک پارامتر می‌توان تخمین‌زننده‌های متعددی را معرفی نمود، لذا بایستی مناسب‌ترین تخمین‌زننده‌ها را انتخاب نماییم. بنابراین، باید شرایط و خواصی را برای یک تخمین‌زننده خوب تعریف کنیم تا بر اساس چنین معیارهایی بتوان در مورد میزان خوبی یک تخمین‌زننده داوری نمود.



بدون تورش یا نااریب بودن[1]

اگر امید ریاضی یک تخمین‌زننده برابر با پارامتر مورد نظر باشد، آن را تخمین‌زنندهٔ بدون تورش (نااریب) گویند:

$$E(\hat{\theta})=\theta \qquad (۱-۶)$$

به‌عبارت دیگر خاصیت نااریبی به‌صورت $E(\hat{\theta})-\theta=0$ است و اگر تخمین‌زننده دارای اریب باشد، این رابطه به‌صورت $E(\hat{\theta})-\theta=B$ می‌باشد که B مقدار تورش (اریب) را نشان می‌دهد.

مثال ۸-۱: فرض کنید X دارای توزیع پوآسن با پارامتر λ باشد. برای تخمین λ از تابع نمونه‌ای یا آماره $\hat{\lambda}=\overline{X}=\frac{\sum x_i}{n}$ استفاده می‌کنیم. می‌توان ثابت نمود که $\overline{X}$ یک تخمین‌زنندهٔ نااریب برای λ است:

$$E(\hat{\lambda})=E(\overline{X})=E\left(\frac{\sum(X_i)}{n}\right)=\frac{\sum E(X)}{n}=\frac{\sum\lambda}{n}=\frac{n\lambda}{n}=\lambda$$

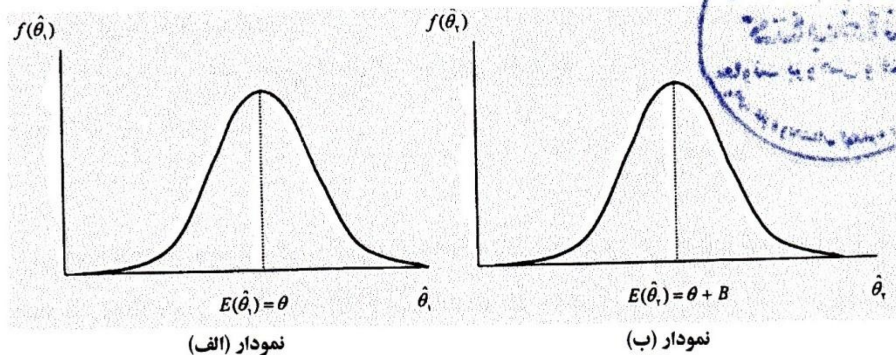
مفهوم بدون تورش بدین معنا است که تخمین‌زننده «به‌طور متوسط» به پارامتر مورد نظر گرایش دارد. در چنین حالتی، مرکز توزیع $\hat{\theta}$ درست بر روی θ قرار دارد و لذا مقدار حاصل از $\hat{\theta}$ به‌طور متوسط به θ نزدیک خواهد بود. اگر دو تخمین‌زننده $\hat{\theta}_1$ و $\hat{\theta}_2$ برای θ داشته باشیم، نمودار الف تخمین بدون تورش را نشان می‌دهد که مقدار $\hat{\theta}_1$ به‌طور متوسط برابر با θ است، ولی نمودار ب تخمین با تورش را نشان می‌دهد که مقدار $\hat{\theta}_2$ به‌طور متوسط برابر با $\theta+B$ است. لذا $\hat{\theta}_2$ مقدار θ را به‌طور متوسط به اندازه B بیش از حد و یا کمتر از حد، برآورد می‌کند.

از طرف دیگر برای تخمین‌زننده‌ای که با افزایش حجم نمونه، مقدار تورش آن کوچک شده و به صفر میل کند، آن را «تخمین‌زنندهٔ بدون تورش حدی» می‌گویند.

$$\lim_{n\to\infty} E(\hat{\theta})=\theta \quad \text{یا} \quad \lim_{n\to\infty} B=0 \qquad (۱-۷)$$

---

[1] unbiased



نمودار (الف) نمودار (ب)

نمودار ۱-۱: تخمین‌زننده‌های بدون تورش و باتورش

مثال ۹-۱: برای تخمین واریانس در جامعه‌ای با توزیع نرمال، نمونه‌ای تصادفی به حجم n انتخاب می‌شود. تخمین‌زننده واریانس را به‌صورت $\tilde{S}^2=\frac{(x_i-\overline{X})^2}{n}$ معرفی می‌کنیم. آیا $\tilde{S}^2$ یک تخمین‌زننده نااریب است؟

بدین منظور از رابطه $\tilde{S}^2$ و توزیع کای‌دو (χ^2) استفاده می‌کنیم:[1]

$$\tilde{S}^2=\frac{\sum(X_i-\overline{X})^2}{\sigma^2}\frac{\sigma^2}{n}=\frac{\sigma^2}{n}\chi^2_{n-1}$$

$$E(\tilde{S}^2)=E(\frac{\sigma^2}{n}\chi^2_{n-1})=\frac{\sigma^2}{n}E(\chi^2_{n-1})=\frac{\sigma^2}{n}(n-1)=\frac{n-1}{n}\sigma^2=\sigma^2-\frac{\sigma^2}{n}$$

مقـــدار

$$E(\tilde{S}^2)-\sigma^2=-\frac{\sigma^2}{n}$$

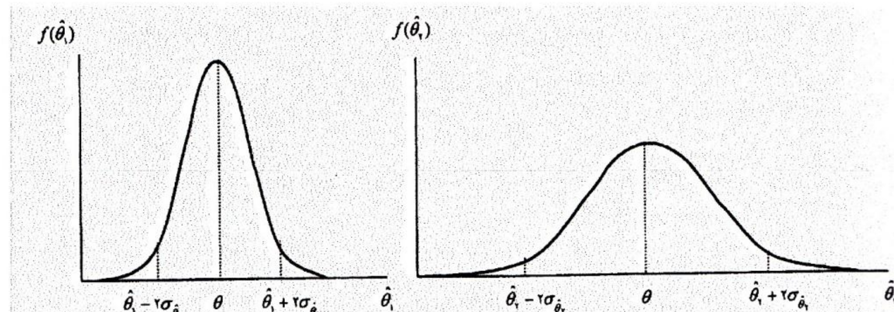
تورش $\tilde{S}^2$ برابر با $-\frac{\sigma^2}{n}$ است. به‌عبارت دیگر $\tilde{S}^2$ به‌طور متوسط مقدار σ^2 را به اندازه $\frac{\sigma^2}{n}$ کمتر از «حدِ واقعی» برآورد می‌کند. اما چون $\lim_{n\to\infty}\frac{\sigma^2}{n}=0$ است، لذا بدون اریب مجانبی می‌باشد.

کارایی (حداقل واریانس)

یکی دیگر از ویژگی‌های تخمین‌زننده‌ها، کوچک بودن واریانس است. برای توضیح این مطلب، فرض کنید برای تخمین پارامتر θ دو تخمین‌زننده $\hat{\theta}_1$ و $\hat{\theta}_2$ داریم که هر دو بدون تورش، ولی واریانس $\hat{\theta}_1$ کمتر از $\hat{\theta}_2$ باشد.

---

[1] توزیع کای‌دو در بخش ۲-۷-۱ ارائه شده است.



نمودار ۲-۱: کارایی تخمین‌زننده‌ها

نمودار فوق نشان می‌دهد که $\hat{\theta}_1$ دارای واریانس کمتری است، لذا تمامی مقادیر ممکن آن در مقایسه با $\hat{\theta}_2$ به θ نزدیکتر هستند. هر تخمین‌زننده‌ای که دارای واریانس کمتری باشد، دارای کارایی بیشتری است. در اینجا کارایی $\hat{\theta}_2$ در مقایسه با $\hat{\theta}_1$ برابر است با:

$$\text{ضریب کارایی } \hat{\theta}_2 \text{ در مقایسه با } \hat{\theta}_1 = \frac{\text{var}(\hat{\theta}_1)}{\text{var}(\hat{\theta}_2)} \qquad (۱-۸)$$

به‌عنوان مثال، اگر واریانس $\hat{\theta}_2$ برابر با ۵ و واریانس $\hat{\theta}_1$ برابر با ۳ باشد، در این صورت کارایی $\hat{\theta}_2$ در مقایسه با $\hat{\theta}_1$ برابر با $0/6 (=\frac{3}{5})$ یا ۶۰ درصد است. این بدان معنا است که کارایی $\hat{\theta}_2$ در نمونه‌ای به حجم ۱۰۰ مشابه کارایی $\hat{\theta}_1$ در نمونه‌ای به حجم ۶۰ می‌باشد.

حال فرض کنید که از بین تمامی تخمین‌زننده‌های θ، تخمین‌زننده‌ای مانند $\hat{\theta}_*$ وجود داشته باشد که دارای کمترین واریانس باشد. بدیهی است که $\hat{\theta}_*$ کارآترین تخمین‌زننده خواهد بود. حال سایر تخمین‌زننده‌های θ را با $\hat{\theta}_*$ مقایسه می‌کنیم. بنابراین ضریب کارایی هر تخمین‌زنندهٔ دلخواه مانند $\hat{\theta}_i$ برابر است با:

$$\text{ضریب کارایی } \hat{\theta}_i = \frac{\text{var}(\hat{\theta}_*)}{\text{var}(\hat{\theta}_i)} \leq 1 \qquad (۱-۹)$$

نامساوی کرامر- رائو، حد پایین واریانس تخمین‌زننده‌های θ است:

$$\text{var}(\hat{\theta}) \geq \frac{1}{E[\frac{\partial}{\partial\theta}\ln L(\theta)]^2} \qquad (۱-۱۰)$$

واریانس کاراترین تخمین زننده برابر با حد پایین این نامساوی است.

مثال ۱۰-۱: برای تخمین پارامتر μ در جامعه‌ای بـا توزیـع نرمـال، نمونـه‌ای بـه حجـم n انتخاب می‌شود. ضریب کارایی $\overline{X}$ به‌عنوان تخمین‌زننده μ را حساب کنید.

$$L(\mu,\sigma^{۲})=\frac{۱}{\left(\sqrt{۲\pi}\sigma\right)^{n}}e^{\frac{\sum_{i=۱}^{n}(x-\mu)^{۲}}{۲\sigma^{۲}}} \Rightarrow \ln L=-n\ln\sqrt{۲\pi}\sigma-\frac{\sum_{i=۱}^{n}(x_i-\mu)^{۲}}{۲\sigma^{۲}}$$

$$\frac{\partial \ln L}{\partial\mu}=\frac{\sum_{i=۱}^{n}(x_i-\mu)}{\sigma^{۲}}\Rightarrow E\left[\frac{\partial \ln L}{\partial\mu}\right]=E\left[\frac{\sum_{i=۱}^{n}(x_i-\mu)}{\sigma^{۲}}\right]^{۲}$$

$$=\frac{\sum_{i=۱}^{n}E(x_i-\mu)^{۲}}{\sigma^{۲}}=\frac{\sum_{i=۱}^{n}\sigma^{۲}}{\sigma^{۴}}=\frac{n}{\sigma^{۲}}$$

توجه شود که در محاسبه امید ریاضی $\left(\sum_{i=۱}^{n}(x_i-\mu)\right)^{۲}$، امید ریاضی حاصل‌ضرب جمـلات متقاطع $\left(\sum_i\sum_j(X_i-\mu)(X_j-\mu)\right)$ برابر صفر است، زیرا X_i ها مـستقل‌انـد. بـدین ترتیـب، کمترین واریانس برای تخمین‌زننده‌های میانگین در یک جامعه نرمال برابر با $\frac{\sigma^{۲}}{n}$ است.

سازگاری

سازگاری یکی از خواص مجانبی تخمین‌زننده‌ها است و بدین معناست کـه اگـر حجـم نمونـه افزایش یابد، احتمال اینکه تفاضل $\hat{\theta}$ از θ از هـر عـدد حقیقـی ماننـد ε کمتـر باشـد، برابـر بـا ۱ خواهد بود.

$$\lim_{n\to\infty}P(|\hat{\theta}-\theta|<\varepsilon)=۱ \qquad (۱-۱۱)$$

این خاصیت بیانگر این است که وقتی حجم نمونه افزایش یابد، مقدار $\hat{\theta}$ کـاملاً بـه θ نزدیـک خواهد شد. شرط فوق را اصـطلاحاً «همگرایـی در احتمـال» مـی‌گوینـد و بـا علامـت plim نـشان می‌دهند:

$$\text{plim}\,\hat{\theta}=\theta \qquad (۱-۱۲)$$

معادل این را می‌توان برحسب واریـانس بیـان نمـود، زیـرا واریـانس نیـز بیـانگر انحـراف $\hat{\theta}$ از $E(\hat{\theta})$ است. اگر تخمین زننده بدون تورش حدی باشد آنگاه شرایط زیر را تأمین می‌کند:

$$\lim_{n\to\infty} E(\hat{\theta})=\theta \quad , \quad \lim_{n\to\infty} \text{var}(\hat{\theta})=۰ \qquad (۱-۱۳)$$

اگر با افزایش حجم نمونه، واریانس $\hat{\theta}$ به سمت صفر میل کند، بدین معناست که انحراف $\hat{\theta}$ از θ به صفر خواهد رسید.

مثال ۱۱-۱: برای تخمین میانگین جامعه نرمال، $\bar{X}$ یک تخمین‌زنندهٔ سازگار است.

$$\text{var}(\bar{X})=\text{var}\left(\frac{\sum X_i}{n}\right)=\frac{\sum \text{var}(X_i)}{n^۲}=\frac{\sum \sigma^۲}{n^۲}=\frac{\sigma^۲}{n}$$

$$\lim_{n\to\infty} \text{var}(\bar{X})=\lim_{n\to\infty}\frac{\sigma^۲}{n}=۰$$

مثال ۱۲-۱: برای تخمین واریانس جامعه نرمال، $\tilde{S}^۲$ یک تخمین‌زنندهٔ باتورش اما سازگار است.

$$\tilde{S}^۲=\frac{\sum (X_i-\bar{x})^۲}{n}=\frac{\sigma^۲}{n}\chi^۲_{n-۱}$$

$$\text{var}(\tilde{S}^۲)=\text{var}(\frac{\sigma^۲}{n}\chi^۲_{n-۱})=\frac{\sigma^۴}{n^۲}\text{var}(\chi^۲_{n-۱})=\frac{\sigma^۴}{n^۲}۲(n-۱)=۲\sigma^۴(\frac{۱}{n}-\frac{۱}{n^۲})$$

$$\lim_{n\to\infty} \text{var}(\tilde{S}^۲)=\lim_{n\to\infty} ۲\sigma^۴(\frac{۱}{n}-\frac{۱}{n^۲})=۰$$

کفایت

$\hat{\theta}$ را یک تخمین‌زنندهٔ کافی برای θ گویند اگر $\hat{\theta}$ از تمام اطلاعات نمونه که مربوط به برآورد θ است استفاده کند. این تعریف معادل است با اینکه توزیع مشترک X_i ها (اطلاعات نمونه) به شرط $\hat{\theta}$ مستقل از θ باشد. توزیع مشترک X_i ها به شرط $\hat{\theta}$ که با $f(x_۱,...,x_n|\hat{\theta})$ نشان داده می‌شود بایستی مستقل از θ باشد. اگر این احتمال بستگی به θ داشته باشد، آنگاه مشاهده یک نمونهٔ خاص بستگی به θ دارد. حال اگر این احتمال مستقل از θ باشد، آنگاه نمونه خاصی مانند $(x_۱,...,x_n)$ که منجر به برآورد θ می‌شوند، برای هر مقدار از θ احتمال یکسانی با سایر نمونه‌ها دارد. احتمال مذکور را می‌توان به‌صورت زیر نوشت:

$$f(x_۱,...,x_n|\hat{\theta})=\frac{f(x_۱,...,x_n,\hat{\theta})}{f(\hat{\theta})} \qquad (۱-۱۴)$$

مثال۱۳-۱: فرض کنید X توزیع دو نقطه‌ای به‌صورت $P(x)=p^{x}(1-p)^{1-x}$ دارد ($x=0,1$). تابع احتمال مشترک X_i ها (تابع درستنمایی) عبارت است از:

$$f(x_1,\dots,x_n,p)=f(x_1|p)\dots f(x_n|p)$$
$$=p^{x_1}(1-p)^{1-x_1}\dots p^{x_n}(1-p)^{1-x_n}=p^{\sum x_i}(1-p)^{n-\sum x_i}$$
$$=p^{y}(1-p)^{n-y} \quad ; \quad y=\sum_{i=1}^{n}x_i$$

از طرف دیگر می‌دانیم که $\hat{p}=\frac{y}{n}$ توزیع دوجمله‌ای دارد.

$$P(\hat{p})=p(\frac{y}{n})=P(y)=C_n^y p^y(1-p)^{n-y}$$

حال $f(x_1,\dots,x_n|\hat{p})$ را می‌نویسیم:

$$f(x_1,\dots,x_n|\hat{p})=\frac{f(x_1,\dots,x_n,p)}{f(\hat{p})}=\frac{p^y(1-p)^{n-y}}{C_n^y p^y(1-p)^{n-y}}$$
$$=\frac{1}{C_n^y}=\frac{1}{C_n^{\sum x_i}}$$

که بستگی به p ندارد. لذا $\hat{p}=\frac{y}{n}$ یک تخمین‌زننده کافی برای p است.

از طرف دیگر می‌توان شرط (۱۴-۱) را برای $\hat{\theta}$ به‌صورت زیر نوشت:

$$f(x_1,\dots,x_n,\theta)=g(\hat{\theta},\theta)h(x_1,\dots,x_n) \qquad (1-15)$$

سمت چپ همان تابع درستنمایی است. در سمت راست، g تابعی از $\hat{\theta}$ و θ است ولی h فقط تابعی از اطلاعات نمونه است.

مثال ۱۴-۱: X توزیع نرمال با میانگین μ و واریانس σ^2 دارد. با فرض معلوم بودن σ^2، برای برآورد μ نمونه‌ای با حجم n انتخاب می‌کنیم. نشان دهید که $\overline{X}$ تخمین‌زننده کافی برای μ است. ابتدا تابع درستنمایی را تشکیل می‌دهیم:

$$f(x_1,\dots,x_n,\mu)=\frac{1}{(\sqrt{2\pi})^n\sigma^n}e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$$

حال تابع درستنمایی را تجزیه می‌کنیم. بدین‌منظور ابتدا عبارت $\sum(x_i-\mu)^2$ را تجزیه می‌کنیم:

$$\sum(x_i-\mu)^2=\sum[(x_i-\bar{x})+(\bar{x}-\mu)]=\sum(x_i-\bar{x})^2+n(\bar{x}-\mu)^2$$

زیرا $\sum(\bar{x}-\mu)^2=n(\bar{x}-\mu)^2$ و $\sum[(x_i-\bar{x})(\bar{x}-\mu)]=0$ است. طبق نتیجهٔ فوق، تابع درستنمایی را به‌صورت زیر می‌نویسیم:

$$f(x_1,\dots,x_n,\mu)=\left[\frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}}e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}}\right]\left[\frac{1}{(\sqrt{2\pi})^{n-1}\sqrt{n}\sigma^{n-1}}e^{-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}}\right]$$

در سمت راست، کروشه اول تابعی از $\bar{X}$ و μ است که همان $g(\bar{X},\mu)$ می‌باشد و کروشه دوم فقط تابعی از X_i ها (اطلاعات نمونه) است که تابعی از μ نیست. بنابراین، $\bar{X}$ تخمین‌زنندهٔ کافی برای μ است.

۱-۵-٤ تخمین فاصله‌ای

آنچه تاکنون بحث شد معروف به «تخمین‌نقطه‌ای» است. این روش، راجع به اینکه تخمین به‌دست آمده تا چه اندازه قابل اطمینان است، بحثی نمی‌کند. در «تخمین فاصله‌ای» برای هر پارامتر، دامنه‌ای به‌دست می‌آید که همراه با یک احتمال است. این احتمال بیانگر سطح اطمینانی است که به فاصله مذکور داریم. بدین لحاظ به آن «فاصله اطمینان» گفته می‌شود. مثلاً اگر با احتمال ۹۵ درصد این فاصله را بسازیم به آن «فاصله اطمینان ۹۵ درصدی» می‌گویند.

برای به‌دست آوردن تخمین فاصله‌ای برای پارامتر θ باتوجه به اینکه $\hat{\theta}$ تخمین‌زنندهٔ θ است، احتمال زیر را تعریف می‌کنیم:

$$P(|\hat{\theta}-\theta|<\varepsilon)=1-\alpha \qquad (1-16)$$

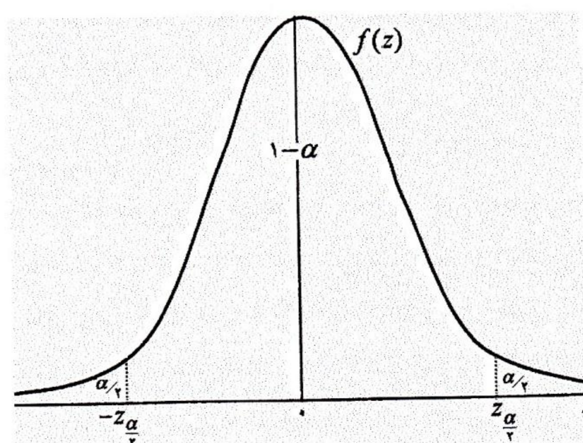
$1-\alpha$ بیانگر سطح اطمینان می‌باشد. احتمال فوق را به‌صورت زیر می‌نویسیم:

$$P(-\varepsilon<\hat{\theta}-\theta<\varepsilon)=1-\alpha$$
$$P(\hat{\theta}-\varepsilon<\theta<\hat{\theta}+\varepsilon)=1-\alpha \qquad (1-17)$$

بنابراین فاصلهٔ $(\hat{\theta}-\varepsilon,\hat{\theta}+\varepsilon)$ با احتمال $1-\alpha$، پارامتر θ را در بر خواهد گرفت. توجه شود که فاصله $(\hat{\theta}-\varepsilon,\hat{\theta}+\varepsilon)$ یا به‌طور خلاصه $\hat{\theta}\pm\varepsilon$ را «تخمین‌زننده فاصله‌ای» می‌گوییم. زیرا $\hat{\theta}$ تابعی از متغیرهای تصادفی است. مقدار حاصل از تخمین‌زنندهٔ فاصله‌ای را «تخمین فاصله‌ای» می‌گوییم.

ε نقش مهمی در تعیین طول فاصله اطمینان دارد. طول فاصله اطمینان برابر با 2ε است، زیرا تفاضل حد بالا $(\hat{\theta}+\varepsilon)$ و حد پایین $(\hat{\theta}-\varepsilon)$ را نشان می‌دهد. اما برای تعیین ε الزاماً باید رابطه $\hat{\theta}$ را با یکی از توزیع‌های شناخته شده پیدا کنیم تا بتوان از جداول آنها استفاده نمود. به‌عنوان مثال یکی از توزیع‌های شناخته شده، توزیع نرمال استاندارد است. برای Z می‌توان احتمال زیر را تعریف نمود:

$$P(-Z_{\frac{\alpha}{2}} \le Z \le Z_{\frac{\alpha}{2}})=1-\alpha \quad \text{یا} \quad P(|Z_{\overline{X}}|<Z_{\frac{\alpha}{2}})=1-\alpha \qquad (1-18)$$



نمودار ۳-۱: توزیع نرمال و تخمین فاصله‌ای

به‌عنوان مثال، اگر بخواهیم برای میانگین جامعه نرمال (μ) تخمین فاصله‌ای بسازیم، از $\overline{X}$ استفاده می‌کنیم که استاندارد شدهٔ آن $Z=(\overline{X}-\mu)\Big/(\frac{\sigma}{\sqrt{n}})$ است. به‌جای Z قرار داده و آن را ساده می‌کنیم:

$$P\left(-Z_{\frac{\alpha}{2}}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<Z_{\frac{\alpha}{2}}\right)=1-\alpha \Rightarrow P\left(-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}<\overline{X}-\mu<Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

بنابراین، $\varepsilon=Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ است و تخمین فاصله‌ای برابر است با:

$$P\left(\overline{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

و یا

$$\left(\overline{X}-Z_{\frac{\alpha}{۲}}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X}+Z_{\frac{\alpha}{۲}}\frac{\sigma}{\sqrt{n}}\right)$$

۱-٦ آزمون فرضیه

آزمون فرضیه در مورد توزیع متغیر تصادفی و پارامترهای آن بحث می‌کند. این آزمون ممکن است راجع به نوع توزیع (مثلاً نرمال بودن و..) باشد یا راجع به پارامترهای این توزیع و یا مقایسه برخی از ویژگی‌های دو یا چند متغیر تصادفی باشد. هر آزمون فرضیه‌ای دارای سه مرحله است:

۱- طرح فرضیه: فرضیه‌های آماری معمولاً به‌صورت دو فرضیه مطرح می‌شوند: یکی فرضیه صفر ($H_۰$) و دیگری فرضیه رقیب ($H_۱$) می‌باشد. به‌عنوان مثال فرضیه صفر بیانگر آن است که نرخ بیکاری (p) برابر با ۰/۲ است، اما فرضیه رقیب می‌گوید که نرخ بیکاری ۰/۱۵ می‌باشد. این فرضیه‌ها به‌صورت زیر نوشته می‌شوند:

$$H_۰: p=۰/۲$$
$$H_۱: p=۰/۱۵$$

اما می‌توان هر یک از حالات زیر را به‌عنوان فرضیه صفر و رقیب در نظر گرفت:

$H_۰: p\leq ۰/۲$	$H_۰: p\geq ۰/۲$	$H_۰: p=۰/۲$	$H_۰: p=۰/۲$	$H_۰: p=۰/۲$
$H_۱: p>۰/۲$	$H_۱: p<۰/۲$	$H_۱: p\neq ۰/۲$	$H_۱: p<۰/۲$	$H_۱: p>۰/۲$

معمولاً در طرح فرضیه‌ها، کل مقادیر مربوط به یک پارامتر را به دو قسمت تقسیم می‌کنند که یک قسمت را $H_۰$ و قسمت دیگر را $H_۱$ پوشش می‌دهد.

۲- معیار آزمون یا آماره آزمون: در این مرحله، قواعد و معیارهایی به‌کار گرفته می‌شود که بر اساس آن بتوان راجع به رد یا عدم رد فرضیه صفر($H_۰$) تصمیم‌گیری نمود. انتخاب این معیارها معادل با معرفی یک آماره است که بتواند اطلاعاتی در خصوص فرضیه‌ها رائه نماید. در این خصوص دو نکته را باید در نظر داشت: اولاً این آماره باید با فرضیه مورد نظر ارتباط داشته باشد. مثلاً اگر فرضیه ما راجع به میانگین جامعه است بایستی از میانگین نمونه ($\overline{X}$) و اگر راجع‌به

واریانس است بایستی از واریانس نمونه ($S^۲$) استفاده نمود. ثانیاً آماره معرفی شده بایستی با یکی از توزیعهای معروف (Z، t، $\chi^۲$ و F) رابطه داشته باشد تا بتوان از جداول آنها استفاده نمود.[1]

۳- قاعده‌ی تصمیم‌گیری: قاعده تصمیم‌گیری را می‌توان بدین صورت مطرح کرد که اگر آماره معرفی شده به اندازه کافی نزدیک به مقدار تعریف شده در فرضیه $H_۰$ باشد، در این صورت $H_۰$ را رد نمی‌کنیم. اما اگر به اندازه کافی از $H_۰$ دور باشد، آن را رد می‌کنیم. بدین منظور بایستی یک عدد بحرانی تعیین کنیم که بتواند معیاری برای نزدیکی یا دوری از فرضیه $H_۰$ باشد.

مثال ۱۵-۱: فرض کنید از جامعه ای نرمال (مثلاً نمرات درس آمار دانشجویان) نمونه‌ای به حجم n انتخاب نموده ایم و می‌خواهیم راجع به میانگین این جامعه (μ) قضاوت کنیم. این آزمون را در سه مرحله انجام می‌دهیم:

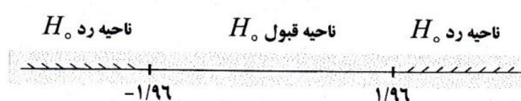
۱- فرضیه را مطرح می کنیم (فرضیه صفر این است که میانگین نمرات برابر با ۱۴ است):

$$H_۰ : \mu = ۱۴$$

$$H_۱ : \mu \neq ۱۴$$

۲- برای قضاوت راجع به این فرضیه میانگین نمونه ($\overline{X}$) را معرفی می‌کنیم. از طرف دیگر می‌دانیم که $\overline{X}$ توزیع نرمال با میانگین μ و واریانس $\frac{\sigma^۲}{n}$ دارد. بنابراین اگر $\overline{X}$ را استاندارد کنیم به توزیع نرمال استاندارد $Z=(\overline{X}-\mu)/(\sigma^۲/n)$ می‌رسیم. اما اگر $\sigma^۲$ مجهول باشد، توزیع t به‌دست می‌آید.

۳- برای تعیین عدد بحرانی یا ناحیه بحرانی از جدول Z یا t استفاده می‌کنیم. به‌عنوان مثال اگر با احتمال ۰/۰۵ بخواهیم این فرضیه را آزمون کنیم آنگاه $Z_{۰/۰۲۵}=۱/۹۶$ است. نحوهٔ تصمیم‌گیری به‌صورت زیر است:



خطای نوع اول و نوع دوم

طبق بحث فوق، α درصد احتمال دارد که آماره مورد نظر در ناحیه بحرانی قرار گیرد، ولی می‌تواند همراه با «درست بودن یا غلط بودن $H_۰$» باشد. اگر $H_۰$ درست باشد و آماره مورد نظر در ناحیه بحرانی قرار گیرد و آن را رد کنیم، آنگاه دچار یک خطا شده‌ایم که معروف به «خطای نوع

[1] این توزیع‌ها در بخش ۷-۱ معرفی شده‌اند.

اول» است. به‌طور کلی در آزمون فرضیه، دو نوع خطا وجود دارد که معروف به خطای نوع اول و خطای نوع دوم می‌باشند.

خطای نوع اول: رد H_0 درست یا قبول H_1 غلط

خطای نوع دوم: رد H_1 درست یا قبول H_0 غلط

۱-۷ مروری بر توزیع‌های مهم

در بخش‌های قبلی، متغیرهای تصادفی، تابع احتمال مشترک نمونه (تابع درستنمایی) و تابعی از متغیرهای نمونه را معرفی کردیم. در این بخش برخی از انواع مهم این توابع را معرفی می‌کنیم که نقش مهمی در آزمون فرضیه و سایر استنتاج‌های آماری دارند.

۱-۷-۱ توزیع نرمال استاندارد

اگر متغیر تصادفی X دارای توزیع نرمال با میانگین μ و واریانس σ^2 باشد، آنگاه استاندارد شدهٔ آن را نرمال استاندارد می‌گویند که به‌صورت زیر تعریف می‌شود:

$$Z=\frac{X-\mu}{\sigma}\sim N(0,1) \qquad (۱-۱۹)$$

لذا Z دارای میانگین ۰ و واریانس ۱ بوده و تابع چگالی آن نیز به‌صورت زیر است:

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \quad , \quad -\infty<z<+\infty \qquad (۱-۲۰)$$

برای محاسبه مقادیر توزیع نرمال از جدول توزیع نرمال استاندارد استفاده می‌شود.

۱-۷-۲ توزیع کای‌دو

این توزیع که نقش مهمی در استنتاج آماری دارد از مجموع مجذور نرمال استاندارد به‌دست می‌آید. بدین منظور فرض کنید از جامعه‌ای که متغیر تصادفی X توزیع نرمال با میانگین μ و واریانس σ^2 دارد، نمونه‌ای به حجم n انتخاب می‌شود. در این صورت هریک از متغیرهای تصادفی نمونه دارای توزیع نرمال هستند و استانداردشدهٔ آنها نیز توزیع نرمال خواهد داشت:

$$X_i\sim N(\mu,\sigma^2) \quad i=1,\cdots,n \;\Rightarrow\; Z_i=\frac{X_i-\mu}{\sigma}\sim N(0,1) \qquad (۱-۲۱)$$

اگر μ مجهول باشد به جای آن از میانگین نمونه یعنی $\overline{X}$ استفاده می‌شود.

حال تابع جدیدی را به‌صورت زیر تعریف می‌کنیم:

$$Y = Z_1^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \qquad (1-22)$$

با فرض معلوم بودن μ و σ^2 این تابع دارای توزیع خاصی به‌نام کای‌دو یا توان دوم کای می‌باشد که آن را با χ_n^2 نشان می‌دهند.[1] در اینجا n را درجه آزادی می‌گویند و بیانگر تعداد متغیرهایی است که مقادیر خود را بدون هیچ محدودیتی اختیار می‌کنند.

اگر در جامعه اصلی میانگین مجهول باشد، آنگاه Y را به‌صورت زیر تعریف می‌کنیم:

$$Y = Z_1^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2} \qquad (1-23)$$

این تابع نیز توزیع کای‌دو دارد. اما درجه آزادی آن $n-1$ است زیرا یک محدودیت به‌صورت $\sum(X_i-\overline{X})=0$ وجود دارد که باعث می‌شود یکی از متغیرها نتواند مقدار خود را آزادانه انتخاب نماید. توزیع χ_k^2 با درجه آزادی k دارای امیدریاضی k و واریانس $2k$ می‌باشد.

یکی از نکات مهم در مورد توزیع کای‌دو، رابطه آن با واریانس نمونه است:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1} = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{\sigma^2}{n-1}\chi_{n-1}^2 \Rightarrow \frac{S^2}{\sigma^2} = \frac{\chi_{n-1}^2}{n-1} \qquad (1-24)$$

توزیع χ^2 هم جمع‌پذیر و هم تجزیه‌پذیر است. بدین معنی که حاصل‌جمع دو توزیع کای‌دو با درجه آزادی‌های k_1 و k_2 دارای توزیع کای‌دو با درجه آزادی k_1+k_2 می‌باشد. از طرف دیگر

۱- تابع چگالی χ_k^2 با درجه آزادی k که مقادیر آن را با y نشان می‌دهیم عبارت است از:

$$f_{\chi^2}(y) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{-\frac{y}{2}} \quad ; \quad y>0$$

در واقع توزیع χ^2 حالت خاصی از توزیع گاما است که در آن، $\alpha=\frac{k}{2}$ و $\beta=\frac{1}{2}$ است.

توزیع کای‌دو با درجه آزادی k را می‌توان به دو توزیع کای‌دو با درجه آزادی‌های k_1 و k_2 تجزیه نمود به‌گونه‌ای که $k=k_1+k_2$ باشد.

توزیع χ^2 را می‌توان به‌صورت ماتریسی نیز به‌دست آورد. بدین منظور فرض کنید که $\mathbf{x}$ بردار متغیرهای نرمال باشد:

$$\mathbf{x}=\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad , \quad E(X_i)=\mu_i \quad , \quad \text{var}(X_i)=\sigma_i^2 \quad , \quad \text{cov}(X_i,X_j)=0 \qquad (1-25)$$

شرط $\text{cov}(X_i,X_j)=0$ بیانگر استقلال X_i و X_j است. امید ریاضی و واریانس بردار $\mathbf{x}$ عبارت است از:

$$E(\mathbf{x})=E\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}=\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}=\boldsymbol{\mu} \qquad (1-26)$$

$$\text{var}(\mathbf{x})=E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\right]=E\begin{bmatrix} X_1-\mu_1 \\ \vdots \\ X_n-\mu_n \end{bmatrix}\left[X_1-\mu_1 \quad \cdots \quad X_n-\mu_n\right]$$

$$=\begin{bmatrix} E(X_1-\mu_1)^2 & \cdots & E[(X_1-\mu_1)(X_n-\mu_n)] \\ \vdots & & \vdots \\ E[(X_n-\mu_n)(X_1-\mu_1)] & \cdots & E(X_n-\mu_n)^2 \end{bmatrix}=\begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}=\boldsymbol{\Omega} \qquad (1-27)$$

بنابراین، $\boldsymbol{\Omega}$ ماتریس قطری است که عنصر iام آن برابر با واریانس X_i است.

با استفاده از محاسبات و نتایج فوق، خواهیم داشت:

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu})\sim\chi^2 \qquad (1-28)$$

$\boldsymbol{\Omega}^{-1}$ برابر است با:

$$\boldsymbol{\Omega}^{-1}=\begin{bmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} \qquad (1-29)$$

با جایگذاری به‌جای $\boldsymbol{\Omega}^{-1}$ و $\mathbf{x}-\boldsymbol{\mu}$، خواهیم داشت:

$$(\mathbf{x}-\boldsymbol{\mu})'\Omega^{-1}(\mathbf{x}-\boldsymbol{\mu})=\begin{bmatrix}X_1-\mu_1 & \cdots & X_n-\mu_n\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_1^2} & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & \frac{1}{\sigma_n^2}\end{bmatrix}\begin{bmatrix}X_1-\mu_1\\ \vdots\\ X_n-\mu_n\end{bmatrix}$$

$$=\sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2=\sum_{i=1}^{n}Z_i^2 \qquad (1-30)$$

بنابراین، عبارت فوق برابر با مجموع مجذور نرمال‌های استاندارد است که توزیع χ^2 خواهد داشت.

۳-۷-۱ توزیع t

توزیع t به‌صورت زیر تعریف می‌شود:

$$t_k=\frac{Z}{\sqrt{\frac{\chi_k^2}{k}}} \qquad (1-31)$$

Z نرمال استاندارد و χ_k^2 نیز توزیع کای‌دو با درجه آزادی k می‌باشد.

به‌عنوان مثال اگر از جامعه‌ای نرمال، نمونه‌ای به حجم n انتخاب شود، در این صورت $\overline{X}$ و S^2 عبارتند از:

$$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}\quad,\quad S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$$

حال $\overline{X}$ را استاندارد کرده و S^2 را نیز بر حسب χ^2 می‌نویسیم:

$$Z_{\overline{X}}=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}=\sqrt{n}\frac{\overline{X}-\mu}{\sigma}\quad,\quad \frac{S^2}{\sigma^2}=\frac{\chi_{n-1}^2}{n-1} \qquad (1-32)$$

توجه شود که $E(\overline{X})=\mu$ و $\text{var}(\overline{X})=\frac{\sigma^2}{n}$ می‌باشد.

حال توزیع t را تشکیل می‌دهیم:

$$t_{n-1} = \frac{Z_{\overline{X}}}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \qquad (1-33)$$

با جایگذاری به‌جای $Z_{\overline{X}}$ و $\frac{\chi^2_{n-1}}{n-1}$ و مرتب نمودن آن خواهیم داشت:

$$t_{n-1} = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \sqrt{n}\frac{\overline{X}-\mu}{S} \qquad (1-34)$$

اگر نتیجهٔ به‌دست آمده را با $Z_{\overline{X}} = \sqrt{n}\frac{\overline{X}-\mu}{\sigma}$ مقایسه کنیم مشاهده می‌شود که t شباهت زیادی به Z دارد و تنها تفاوت آنها در این است که به‌جای σ (انحراف معیار جامعه) از S (انحراف معیار نمونه) استفاده شده است.

۱-۷-۴ توزیع F

توزیع F کاربردهای بسیاری در مباحث اقتصاد سنجی، به ویژه در آزمون فرضیه‌ها، مقایسه رگرسیون‌ها و یا مقایسه اثر عوامل مختلف بر یک متغیر دارد. توزیع F از تقسیم دو توزیع کای‌دو به‌دست می‌آید:

$$F_{k_1,k_2} = \frac{\chi^2_{k_1}}{k_1} \Big/ \frac{\chi^2_{k_2}}{k_2} \qquad (1-35)$$

توزیع F دارای دو درجه آزادی می‌باشد که یکی مربوط به‌صورت کسر و دیگری مربوط به مخرج کسر است. با توجه به رابطه χ^2 و S^2 می‌توان توزیع F را به‌صورت نسبت واریانس‌های نمونه نوشت. فرض کنید از دو جامعه، نمونه‌هایی به حجم n_1 و n_2 انتخاب شود. در این صورت طبق رابطه (۱-۳۲)، روابط زیر را داریم:

$$\frac{S_1^2}{\sigma_1^2} = \frac{\chi^2_{n_1-1}}{n_1-1} \quad \text{و} \quad \frac{S_2^2}{\sigma_2^2} = \frac{\chi^2_{n_2-1}}{n_2-1}$$

با جایگذاری در F و مرتب نمودن آن خواهیم داشت:

$$F_{n_1-1,n_2-1} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \qquad (1-36)$$

۱-۷-۵ حالت‌های خاص توزیع‌های χ^2 ، t و F

حالت‌های خاص توزیع‌های χ^2 ، t و F وهمچنین روابط آنها عبارت است از:

۱ ـ اگر درجه آزادی کای‌دو به سمت بی‌نهایت میل کند، خواهیم داشت:

$$\lim_{k\to\infty}\frac{\chi_k^2}{k}=1 \qquad (1-37)$$

۲ ـ اگر درجه آزادی کای‌دو بزرگتر از ۳۰ باشد، می‌توان از توزیع نرمال به‌عنوان تقریب مناسبی برای آن استفاده نمود:

$$\chi_k^2 \underset{k\to\infty}{\sim} N(k,2k) \qquad (1-38)$$

اما چون توزیع کای‌دو به‌کندی به توزیع نرمال گرایش دارد از تبدیل فیشر استفاده می‌شود که طبق آن رابطه زیر برقرار است:

$$\sqrt{2\chi_k^2} \underset{k>30}{\sim} N(\sqrt{2k-1},1)$$

بنابر این، استاندارد شدهٔ آن عبارت است از:

$$Z=\frac{\sqrt{2\chi_k^2}-\sqrt{2k-1}}{1}=\sqrt{2\chi_k^2}-\sqrt{2k-1} \underset{k>30}{\sim} N(0,1) \qquad (1-39)$$

مقادیر توزیع کای‌دو برابر است با:

$$\sqrt{2\chi_k^2}=Z+\sqrt{2k-1} \Rightarrow \chi_k^2=\frac{1}{2}\left(Z+\sqrt{2k-1}\right)^2 \qquad (1-40)$$

به‌عنوان مثال، مقدار ۵ درصدی توزع کای‌دو با درجه آزادی k=۴۰ بابر است با:

$$\chi^2_{0/05,40}=\frac{1}{2}(Z_{0/05}+\sqrt{2(40)-1})^2=\frac{1}{2}(1/64+\sqrt{79})^2=55/42$$

که مقدار دقیق $\chi^2_{0/05,40}$ با استفاده از جدول کای‌دو برابر با ۵۵/۸ می‌باشد.

۳- اگر در توزیع F درجه آزادی مخرج، بی‌نهایت باشد خواهیم داشت:

$$F_{k_1,\infty} = \frac{\chi^2_{k_1}/k_1}{1} = \frac{\chi^2_{k_1}}{k_1} \qquad (1-41)$$

اگر درجه آزادی صورت بی‌نهایت باشد، خواهیم داشت:

$$F_{\infty,k_2} = \frac{1}{\chi^2_{k_2}/k_2} = \frac{k_2}{\chi^2_{k_2}} = \frac{1}{F_{k_2,\infty}} \qquad (1-42)$$

۴- اگر در توزیع F ، درجه آزادی صورت برابر با ۱ باشد، آنگاه توزیع F با مجذور توزیع t برابر است:

$$F_{1,k_2} = \frac{\chi^2_1/1}{\chi^2_{k_2}/k_2} = \frac{Z^2}{\chi^2_{k_2}/k_2} = t^2_{k_2} \qquad (1-43)$$

۸-۱ انواع داده‌ها

داده‌های مورد استفاده در مباحث اقتصادی و مالی را می‌توان به سه دسته کلی تقسیم نمود: سری‌های زمانی، داده‌های مقطعی و داده‌های پانل (ترکیبی).

داده‌های سری زمانی مربوط به یک دوره زمانی معین است که ممکن است سالانه، فصلی، ماهانه، هفتگی، روزانه و یا حتی ساعتی و کمتر باشد.

داده‌های مقطعی بیانگر اطلاعات مربوط به یک متغیر در یک زمان معین می‌باشد. مانند مصرف نان در گروههای درآمدی مختلف در سال ۱۳۸۰، بازده سهام شرکتهای مختلف در سال ۱۳۸۸، تغییرات قیمت سهام شرکتهای مختلف در ۱۳۸۸/۱۰/۲۰ و رشد اقتصادی کشورهای مختلف در سال ۲۰۱۰.

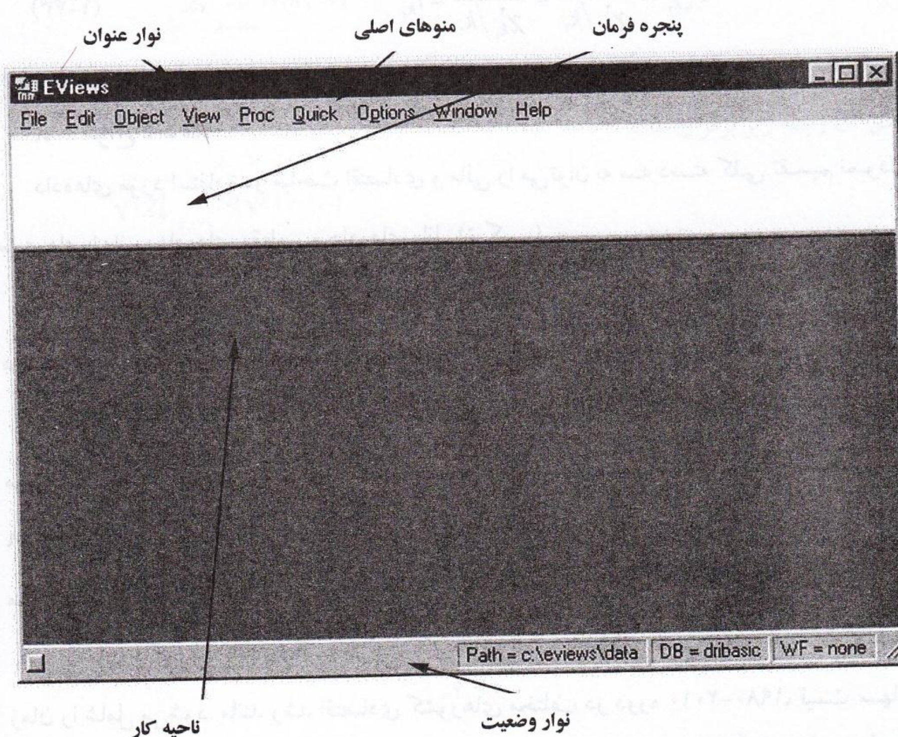
داده‌های ترکیبی یا پانل، ترکیبی از داده‌های سری زمانی و مقطعی است، یعنی هم مقطع و هم زمان را شامل می‌شود. مانند رشد اقتصادی کشورهای مختلف در دوره ۲۰۱۰-۱۹۸۰، قیمت سهام شرکتهای مختلف در دوره ۱۳۸۵-۱۳۸۰. اگر طول دوره را با T و تعداد مقطع‌ها را با n نشان دهیم، کل مشاهدات برابر با nT خواهد بود.

۹-۱ آشنایی مقدماتی با نرم‌افزار Eviews

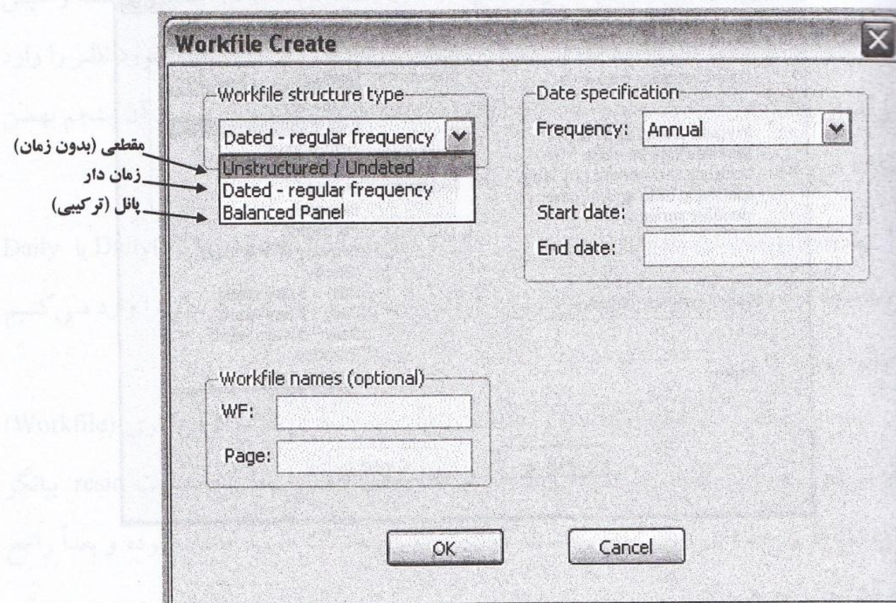
Eviews از جمله نرم‌افزارهایی است که کار با آن بسیار ساده است. در اینجا نکات اولیه جهت کار با Eviews7 را ارائه می‌کنیم. سایر نکات در فصول بعدی همراه با مباحث اقتصاد سنجی ارائه خواهد شد.

ایجاد فایل کاری (Workfile)

بعد از نصب Eviews و اجرای آن، پنجره Eviews باز می‌شود که به شکل زیر بوده و از این به بعد آن را پنجره اصلی می گوییم.



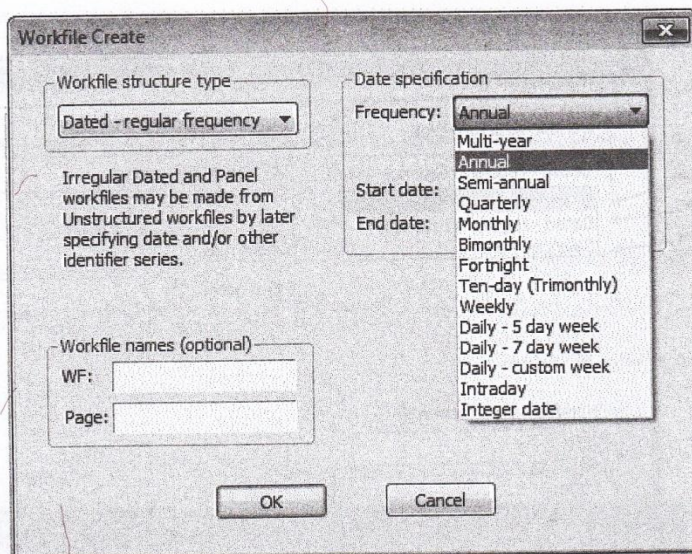
برای ایجاد فایل کاری، ابتدا در پنجره اصلی از طریق منوی File گزینه New را انتخاب می‌کنیم که به دنبال آن پنجره فایل کاری تحت عنوان Workfile Create باز می‌شود.



در سمت چپ این پنجره باید نوع داده‌ها را مشخص کنیم که آیا مقطعی (بدون زمان) یا سری زمانی (زمان‌دار) است و یا پانل (ترکیبی) می‌باشد. در مورد داده‌های ترکیبی در فصول بعدی بحث خواهیم کرد.

اگر داده‌ها به‌صورت مقطعی باشد، ابتدا گزینه Unstructured/Undated را انتخاب کرده و سپس تعداد مشاهدات را در قسمت Observations که در سمت راست نمایان می‌شود، وارد می‌کنیم.

اگر داده‌ها به‌صورت سری زمانی باشد، ابتدا گزینه Dated-regular frequency را انتخاب کرده و سپس در سمت راست، بر حسب نوع داده‌ها یکی از موارد سالانه، فصلی، ماهانه، هفتگی و روزانه و ... را انتخاب می‌کنیم.



برخی از این موارد به‌صورت زیر می‌باشد:

داده‌های سالانه: در سمت راست، گزینه Annual را انتخاب می‌کنیم. اگر سال شروع ۱۳۵۰ و سال پایان ۱۳۸۰ باشد، زمان شروع و پایان را به‌صورت زیر وارد می‌کنیم:

Start date	End date
1350	1380

داده‌های فصلی: در سمت راست، گزینه Quarterly را انتخاب می‌کنیم. اگر زمان شروع، تابستان ۱۳۵۰ و زمان پایان، بهار ۱۳۹۰ باشد، آن را به‌صورت زیر وارد می‌کنیم:

Start date	End date
1350:2	1390:1

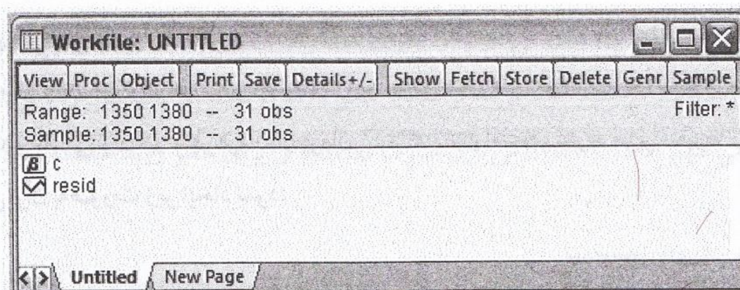
داده‌های ماهانه: در سمت راست، گزینه Monthly را انتخاب می‌کنیم. اگر زمان شروع، بهمن ۱۳۵۰ و زمان پایان، خرداد ۱۳۹۰ باشد، آن را به‌صورت زیر وارد می‌کنیم:

Start date	End date
1350:11	1390:3

داده‌های هفتگی: در سمت راست گزینه Weekly را انتخاب می‌کنیم. در اینجا ابتدا هفته و سپس سال را وارد می‌کنیم. برای وارد نمودن شماره هفته، تاریخ شروع آن در ماه مورد نظر را وارد می‌کنیم (سال: روز: ماه). به‌عنوان مثال 11:5:1380 بیانگر هفته‌ای است که شروع آن پنجم بهمن ۱۳۸۰ می‌باشد.

داده‌های روزانه: برای داده‌های روزانه، ابتدا یکی از گزینه‌های Daily(5 Day Weeks) یا Daily (7 Day Weeks) را انتخاب کرده و سپس به ترتیب ماه، روز و سال مورد نظر را وارد می‌کنیم (سال: روز: ماه).

بعد از مشخص نمودن نوع داده‌ها وانتخاب OK، پنجره دیگری به نام فایل کاری (Workfile) باز می‌شود. در این پنجره، حرف c بیانگر پارامترهای مدل تخمینی است و عبارت resid بیانگر باقیمانده‌ها یا خطاهای مدل تخمینی است که در تخمین معادلات مورد استفاده بوده و بعداً راجع به آن بحث خواهیم کرد.



ورود داده‌ها

برای ورود داده‌ها می‌توان مستقیماً هر متغیری را ایجاد کرده و داده‌های آن را به‌صورت دستی وارد نمود. به‌عنوان مثال برای ورود داده‌های X ابتدا در پنجره فرمان، عبارت data x را نوشته و با زدن کلید Enter پنجره زیر را باز می‌کنیم.

Group: UNTITLED Workfil...

View | Proc | Object | Print | Name | Freeze | Default

obs	obs	X
1345	1345	NA
1346	1346	NA
1347	1347	NA
1348	1348	NA
1349	1349	NA
1350	1350	NA
1351	1351	NA
1352	1352	NA
1353	1353	NA
1354	1354	NA
1355	1355	NA
1356	1356	NA
1357	1357	NA
1358	1358	NA
1359	1359	NA
1360		

در این پنجره در مقابل هر سال داده‌های آن سال را وارد و در پایان، با بستن این پنجره، نام متغیر مورد نظر (x) در فایل کاری نمایان می‌شود. همچنین برای ورود داده‌ها می‌توان آنها را از نرم‌افزارهای دیگری مانند Excel وارد Eviews نمود.

ایجاد متغیرهای جدید

با فرمان genr می‌توان هر متغیر جدیدی را بر اساس متغیرهای موجود، ایجاد نمود. مثلاً اگر z برابر با x^2 باشد می‌توان در سطر فرمان از فرمان genr z=x^2 استفاده نمود. سایر متغیرها را به‌عنوان نمونه می‌توان به‌صورت زیر ایجاد نمود:

جمع u=x+y

تقسیم u=x/5 , u=x/y

لگاریتم u=log(x)

نمایی u=exp(x)

وقفه‌ها و تفاضل‌ها

اگر متغیر x_t بیانگر مقدار x در سال t باشد، در نرم‌افزار Eviews با x نشان داده می‌شود. ولی برای x_{t-1} که مقدار x را در سال در سال t-1 (سال گذشته) و یا مقدار x را با یک وقفه بیان می‌کند،

در Eviews با x(-1) نشان داده می‌شود. همچنین برای دو وقفه (دو سال قبل) از x(-2) استفاده می‌شود. علاوه براین، برای سال‌های بعد از t، مانند t+1 و t+2 از x(1) و x(2) استفاده می‌شود.

طبق تعاریف فوق برای محاسبه تفاضل مرتبه اول (تغییرات مرتبه اول) از x-x(-1) استفاده می‌شود. اگر تفاضل مرتبه اول x را با dx نشان دهیم، در این صورت از فرمان زیر استفاده می‌کنیم:

genr dx=x-x(-1) یا genr dx=d(x)

همچنین برای محاسبه نرخ رشد x ، یعنی gx، از فرمان زیر استفاده می‌کنیم:

genr gx=(x-x(-1))/x(-1)=d(x)/x(-1)

تغییر نام متغیرها

برای تغییر نام متغیرها از فرمان «تغییر نام» استفاده کنیم که علامت اختصاری آن r می‌باشد. مثلاً برای تغییر نام x به y به‌صورت زیر عمل می‌کنیم:

r x y

متغیر زمان

در مواردی نیاز به «متغیر روند» داریم. به‌عنوان مثال اگر داده‌های ما از سال ۱۳۵۱ تا ۱۳۹۰ باشد، متغیر روند که با t نشان می‌دهیم از ۱ برای سال ۱۳۵۱ شروع شده و تا ۴۰ برای سال ۱۳۹۰ ادامه می‌یابد. مانند هر متغیر معمولی، اگر مقادیر متغیر روند را با t نشان دهیم، آن را با فرمان « data t» ایجاد می‌کنیم. اما در Eviews می‌توان عبارت trend@ را برای متغیر روند استفاده نمود.

اصلاح داده‌های وارد شده

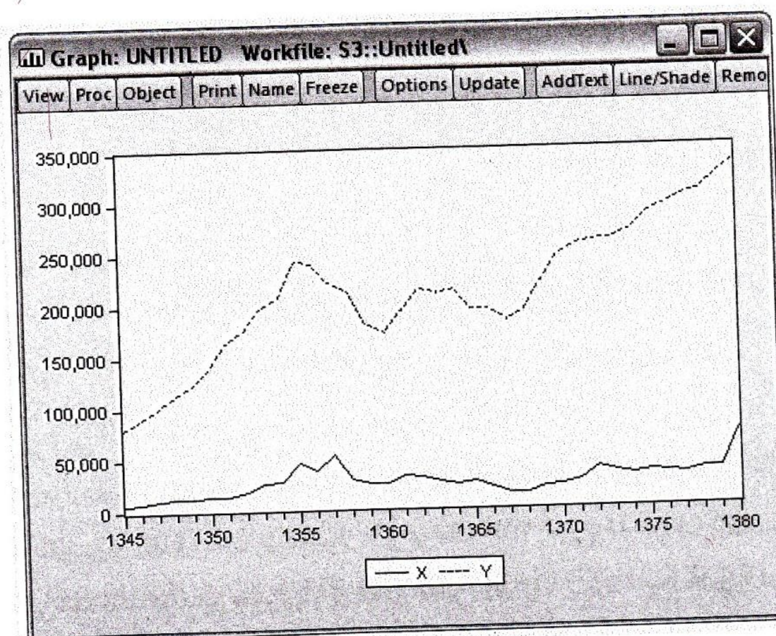
همان طور که قبلاً دیدم برای وارد نمودن داده‌های متغیر x از فرمان «data x» استفاده می‌کنیم. حال اگر در وارد نمودن داده‌ها، اشتباهی رخ داده باشد مجدداً فرمان «data x» را اجرا کرده و اصلاحات مورد نظر را انجام می‌دهیم.

مشاهدهٔ داده‌های واردشده

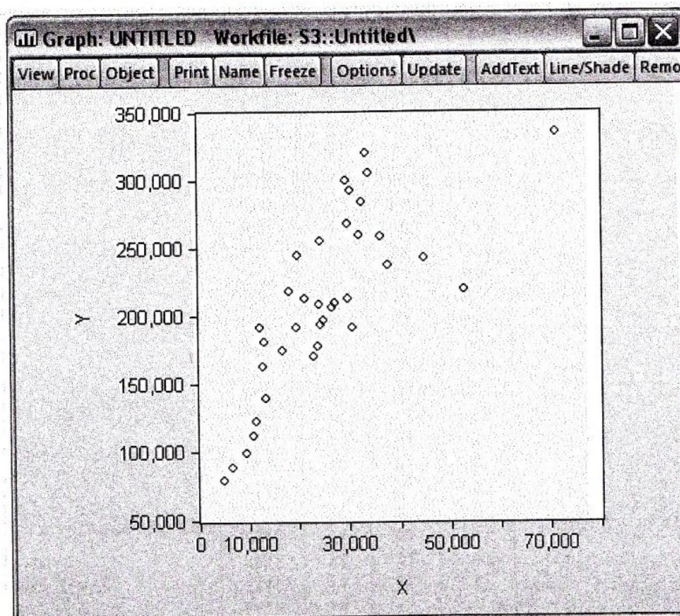
برای مشاهده داده‌های وارد شده از فرمان show استفاده می‌شود. مثلاً برای مشاهده مقادیر x و y از فرمان «show x y» استفاده می‌کنیم.

نمودارها

در Eviews امکان ترسیم نمودارهای مختلف وجود دارد. مثلاً برای ترسیم داده‌های سری زمانی از فرمان plot استفاده می‌شود. به‌عنوان مثال برای ترسیم داده‌های x از فرمان «plot x» وبرای ترسیم همزمان داده‌های x و y از فرمان «plot x y» استفاده می‌کنیم.



در بسیاری از موارد نیاز به ترسیم داده‌های یک متغیر در مقابل دیگری داریم. در این صورت می‌توان از فرمان scat استفاده نمود. برای رسم داده‌های y در مقابل x (y روی محور عمودی و x روی محور افقی) از فرمان «scat x y» استفاده می‌کنیم. با اجرای این فرمان، نمودار پراکندگی ترسیم می‌گردد.



تغییر دوره زمانی

در بالای فایل کاری، عبارت دامنه(Range) و نمونه (Sample) را ملاحظه می‌کنیم. به‌عنوان مثال اگر داده‌های ما شامل دوره ۱۳۵۰ تا ۱۳۸۰ باشد، در این صورت دامنه و نمونه شامل دوره ۱۳۵۰ تا ۱۳۸۰ می‌باشد. اگر بخواهیم دامنه را تا سال ۱۳۹۰ گسترش دهیم در سطر فرمان از فرمان «range 1350 1390» استفاده می‌کنیم. معمولاً نیاز کمتری به تغییر دامنه پیدا می‌کنیم. موارد استفاده از تغییر نمونه بسیار زیاد است. به‌عنوان مثال، اگر بخواهیم فقط با داده‌های دوره ۱۳۵۰ تا ۱۳۷۰ کار کنیم در صورت دوره نمونه را با فرمان «smpl 1350 1370» تغییر می‌دهیم. این فرمان را می‌توان برای هر دوره دیگری نیز به‌کار برد. به‌عنوان نمونه اگر بخواهیم نمونه ما شامل سال ۱۳۵۷ نباشد از فرمان «smpl 1350 1356 1358 1380» استفاده می‌کنیم.

مسائل

۱- یک فایل کاری در Eviews بسازید و داده‌های زیر را وارد کنید.

سال	X	Y	Z
۱۳۵۱	۱	۵۰	-۱۰
۱۳۵۲	۱	۴۸	-۸
۱۳۵۳	۲	۴۶	-۹
۱۳۵۴	۲	۵۰	-۶
۱۳۵۵	۳	۴۲	-۵
۱۳۵۶	۳	۴۴	-۵
۱۳۵۷	۴	۴۰	-۸
۱۳۵۸	۶	۴۰	-۲
۱۳۵۹	۵	۳۸	۱
۱۳۶۰	۸	۳۶	۱
۱۳۶۱	۱۰	۳۵	۲
۱۳۶۲	۱۰	۳۵	۴
۱۳۶۳	۱۲	۳۰	۴
۱۳۶۴	۱۴	۲۸	۵
۱۳۶۵	۱۳	۲۶	۶
۱۳۶۶	۱۶	۲۵	۶
۱۳۶۷	۱۸	۲۵	۷
۱۳۶۸	۱۸	۲۴	۸
۱۳۷۰	۲۰	۲۲	۹
۱۳۷۱	۲۰	۲۰	۱۰
۱۳۷۲	۲۲	۲۲	۱۰
۱۳۷۳	۲۳	۱۸	۱۲
۱۳۷۴	۲۲	۱۸	۱۰
۱۳۷۵	۲۵	۱۶	۱۵
۱۳۷۶	۲۷	۱۶	۱۵
۱۳۷۷	۳۰	۱۴	۱۸
۱۳۷۸	۳۰	۱۲	۲۰
۱۳۷۹	۳۳	۱۲	۲۰
۱۳۸۰	۳۳	۱۰	۲۲

۲- در مسئله ۱، نمودار سری‌های زمانی را برای X، Y و Z ترسیم کنید.

۳- در مسئله ۱، Y را در مقابل Z رسم کنید.

۴- در مسئله ۱، تغییرات Y و Z را حساب کرده و سپس تغییرات Y را در مقابل تغییرات Z رسم کنید. چه تفاوتی با نمودار مسئلهٔ ۳ مشاهده می‌کنید.

۵- در مسئله ۱ نرخ رشد X، Y و Z را حساب کنید و نمودار آنها را ترسیم کنید.

۶- اگر X توزیع نرمال بـا تـابع چگـالی $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ داشـته باشـد، ضـرایب μ و σ^2 را با روش گشتاورها برآورد کنید.

۷- X توزیع هندسی با تابع احتمال $P(x)=p(1-p)^{x-1}$ دارد. ضریب p را بـا روش گـشتاورها برآورد کنید.

۸- X توزیع نرمال با تابع چگالی $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ دارد. ضرایب μ و σ^2 را با روش حداکثر درستنمایی برآورد کنید.

۹- X توزیع هندسی با تابع احتمال $P(x)=p(1-p)^{x-1}$ دارد. ضریب p را با روش حداکثر درستنمایی برآورد کنید.

۱۰- در مسئله ۳، خاصیت بدون تورش بودن تخمین‌زننده‌های حداکثر درستنمایی را بررسی کنید.

۱۱- در مسئله ۳، آیا تخمین‌زننده حداکثر درستنمایی، حد پایین نامساوی رائو-کرامر را تأمین می‌کند.

۱۲- در مسئله ۳، تخمین‌زنندهٔ σ^2 که به‌صورت $\frac{\Sigma(X_i-\bar{X})^2}{n}$ به‌دست می‌آید، آیا در مقایسه با تخمین‌زننده بدون تورش $S^2=\frac{\Sigma(X_i-\bar{X})^2}{n-1}$ کارایی بیشتر دارد.

۱۳- در مسئله ۴، آیا تخمین‌زننده p حد پایین نامساوی رائو-کرامر را تأمین می کند.

۱۴- خاصیت سازگاری را برای تخمین‌زننده‌های μ و σ^2 در مسئله ۸ بررسی کنید.

۱۵- خاصیت سازگاری را برای تخمین‌زننده p در مسئله ۹ بررسی کنید.

۱۶- اگر u_i توزیع نرمال با میانگین صفر و واریانس σ^2 داشته باشد، ثابت کنید که $\frac{\mathbf{u'u}}{\sigma^2}$ توزیع کای‌دو دارد که $\mathbf{u'}=(u_1,u_2,\dots,u_n)$ است.

۱۷- در مسئله ۱۶ ثابت کنید که نسبت $\sqrt{n-1}\frac{\bar{u}}{\sqrt{\mathbf{u'u}}}$ توزیع t با درجه آزادی $n-1$ دارد ($\bar{u}$ میانگین u_i است).

۱۸- ثابت کنید که $F_{1,n}=t_n^2$ است.

۱۹- X توزیع نمایی به‌صورت $f(x)=\frac{1}{\beta}e^{-\frac{x}{\beta}}$ دارد.

الف) با استفاده از روش گشتاورها، تخمین‌زنندهٔ β را به‌دست آورید.

ب) ثابت کنید که تخمین‌زنندهٔ حداکثر درستنمایی برای β برابر با $\bar{X}$ است.

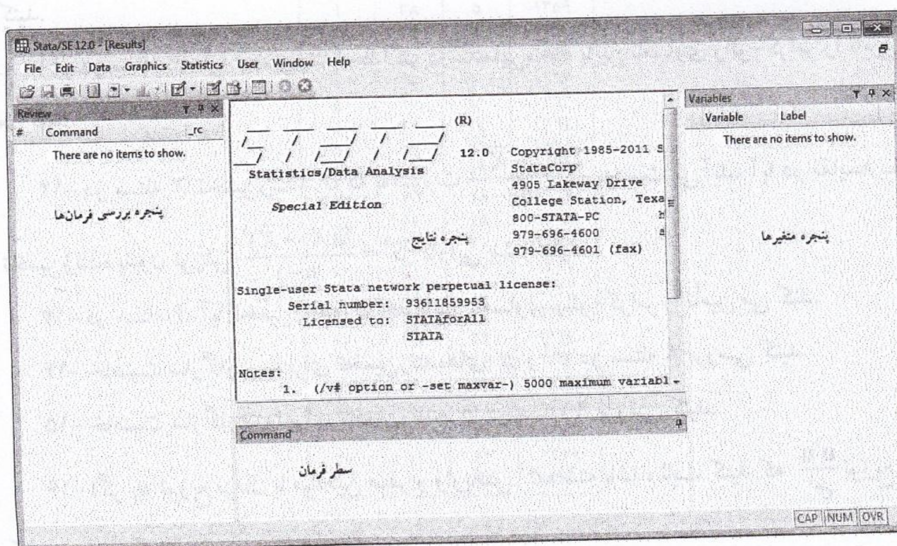
ج) آیا تخمین‌زنندهٔ حداکثر درستنمایی، شرایط بدون تورش، سازگاری، حد پایین نامساوی رائو-کرامر و کفایت را تأمین می‌کند.

ضمیمه فصل ۱: مروری بر نرم‌افزار Stata

معرفی بخش‌های اولیه Stata

بعداز اجرای نرم‌افزار Stata، ابتدا پنجره زیر را مشاهده خواهید کرد که دارای چهار قسمت است:

۱- سطر فرمان که می‌توان فرمان‌های مختلف را در آن تایپ کرد.
۲- پنجره بررسی فرمان‌های قبلی است که فرمان‌های قبلی را نشان می‌دهد.
۳- پنجره متغیرها که نام متغیرهای موجود در فایل داده‌ها را نشان می‌دهد.
٤- پنجره نتایج که نتایج حاصل از اجرای فرمان‌ها را نشان می‌هد.



نوار بالای پنجره فوق، منوهای اصلی را نشان می‌دهد که هر یک از آنها کارکردهای خاص خود را دارد که در قسمت‌های بعدی به آنها اشاره خواهیم کرد.

ایجاد فایل کاری و ورود داده‌ها

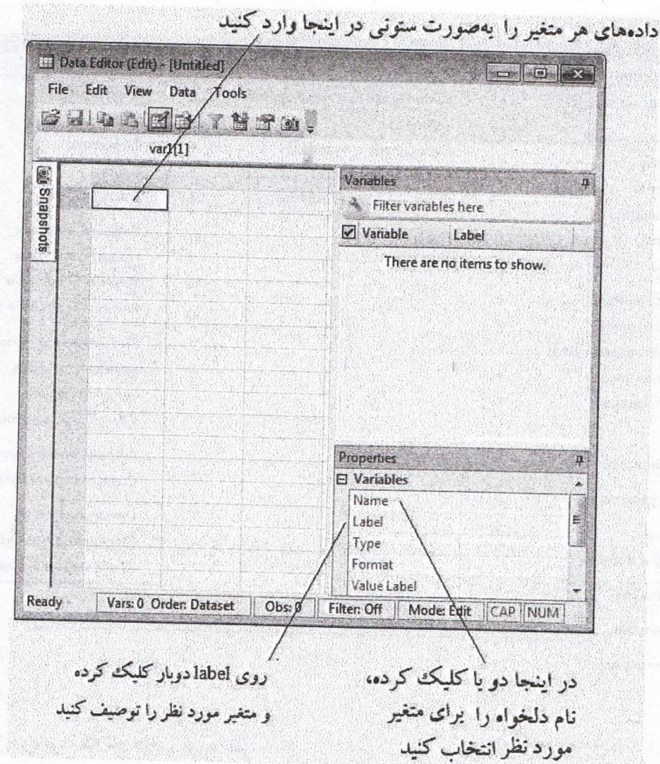
ابتدا بایستی یک فایل کاری برای وارد نمودن داده‌ها ایجاد نماییم. بدین منظور مسیر زیر را انتخاب می‌کنیم:

$$\text{Data} \rightarrow \text{Data editor}$$

پنجره data editor برای ورود داده‌ها باز می‌شود. در این پنجره، هر ستون برای یک متغیر می‌باشد. با کلیک روی سطر اول هر ستون می‌توانید داده‌های متغیر مورد نظر را وارد کنید. داده‌ها را می‌توان به‌صورت دستی یا از برنامه‌های دیگر مانند excel یا eviews و ... کپی نمود.

در این پنجره، delete برای حذف متغیر، hide برای پنهان کردن نام متغیر، restore برای ذخیره تغییرات انجام شده می‌باشد.

در اینجا لازم است متغیر زمان را تعریف کنیم که برای سری‌های زمانی بیانگر زمان یا دورهٔ نمونه است.



علاوه بر شیوه فوق، می‌توان برای تغییر نام متغیره از فرمان زیر استفاده نمود:

rename var1 y یا ren var1 y

اجرای فرمان فوق، نام متغیر var1 را به y تغییر می‌دهد.

مشاهده و اصلاح داده‌های وارد شده

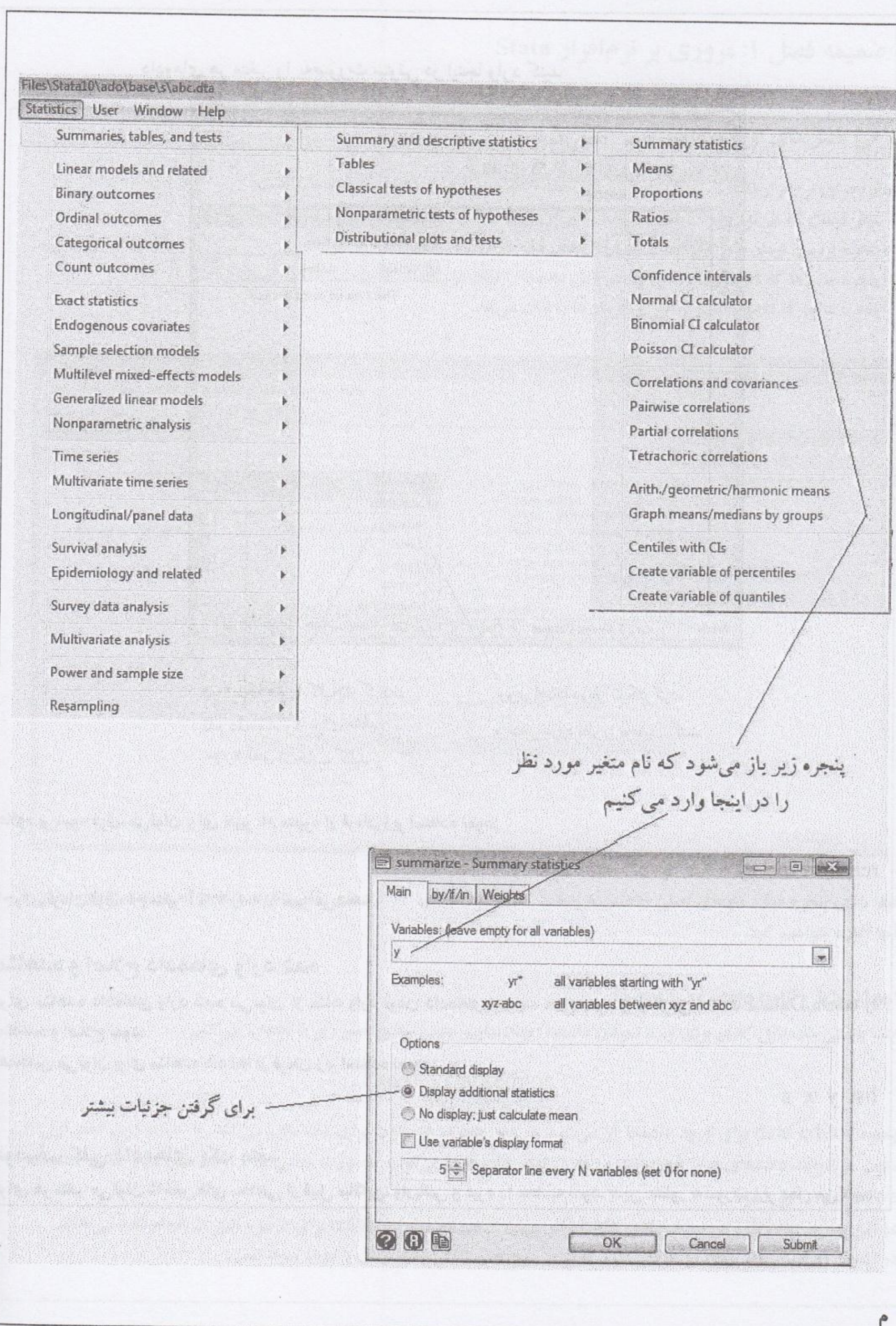
برای مشاهده داده‌های وارد شده می‌توان از مشابه وارد نودن داده‌های جدید، عمل نمود و از طریق Data Editor داده‌ها را مشاهده و اصلاح نمود.

همچنین می‌توان برای مشاهده داده‌ها از فرمان زیر استفاده نمود:

list y x z

توصیف کلی داده‌های یک متغیر

برای هر متغیر می‌توان شاخص‌های مختلفی از قبیل میانگین، واریانس و غیره را محاسبه نمود. بدین منظور به‌صورت زیر عمل می‌کنیم:


Files\Stata10\ado\base\s\abc.dta
Statistics User Window Help
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Exact statistics
Endogenous covariates
Sample selection models
Multilevel mixed-effects models
Generalized linear models
Nonparametric analysis
Time series
Multivariate time series
Longitudinal/panel data
Survival analysis
Epidemiology and related
Survey data analysis
Multivariate analysis
Power and sample size
Resampling
Summary and descriptive statistics
Tables
Classical tests of hypotheses
Nonparametric tests of hypotheses
Distributional plots and tests
Summary statistics
Means
Proportions
Ratios
Totals
Confidence intervals
Normal CI calculator
Binomial CI calculator
Poisson CI calculator
Correlations and covariances
Pairwise correlations
Partial correlations
Tetrachoric correlations
Arith./geometric/harmonic means
Graph means/medians by groups
Centiles with CIs
Create variable of percentiles
Create variable of quantiles
پنجره زیر باز می‌شود که نام متغیر مورد نظر
را در اینجا وارد می‌کنیم
summarize - Summary statistics
Main by/if/in Weights
Variables: (leave empty for all variables)
y
Examples: yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc
Options
Standard display
Display additional statistics
No display; just calculate mean
Use variable's display format
5 Separator line every N variables (set 0 for none)
برای گرفتن جزئیات بیشتر
OK Cancel Submit

مشخصات متغیر y به‌صورت زیر نشان داده می‌شود:

```
                          y
-------------------------------------------------------------
      Percentiles      Smallest
 1%        11             11
 5%        11             11
10%        11.5           12       Obs                  20
25%        13.5           12       Sum of Wgt.          20

50%        17.5                    Mean               18.3
                        Largest    Std. Dev.      5.582869
75%        23.5           24
90%        26             25       Variance       31.16842
95%        27.5           27       Skewness          .2422
99%        28             28       Kurtosis       1.704557
```

صدک‌ها، تعداد مشاهدات، میانگین، انحراف معیار، چارک اول، میانه (چارک دوم)، واریانس، چولگی، کشیدگی، چارک سوم

نتایج فوق را می‌توان با فرمان سادهٔ sum y, detail نیز به‌دست آورد. علاوه بر این، اگر یکی از فرمان‌های زیر را اجرا کنیم، خلاصه‌ای از وضعیت متغیر مورد نظر ارائه می‌شود:

Summarize ip یا summ ip یا sum ip

که نتایج به‌صورت زیر نشان داده شده می‌شود:

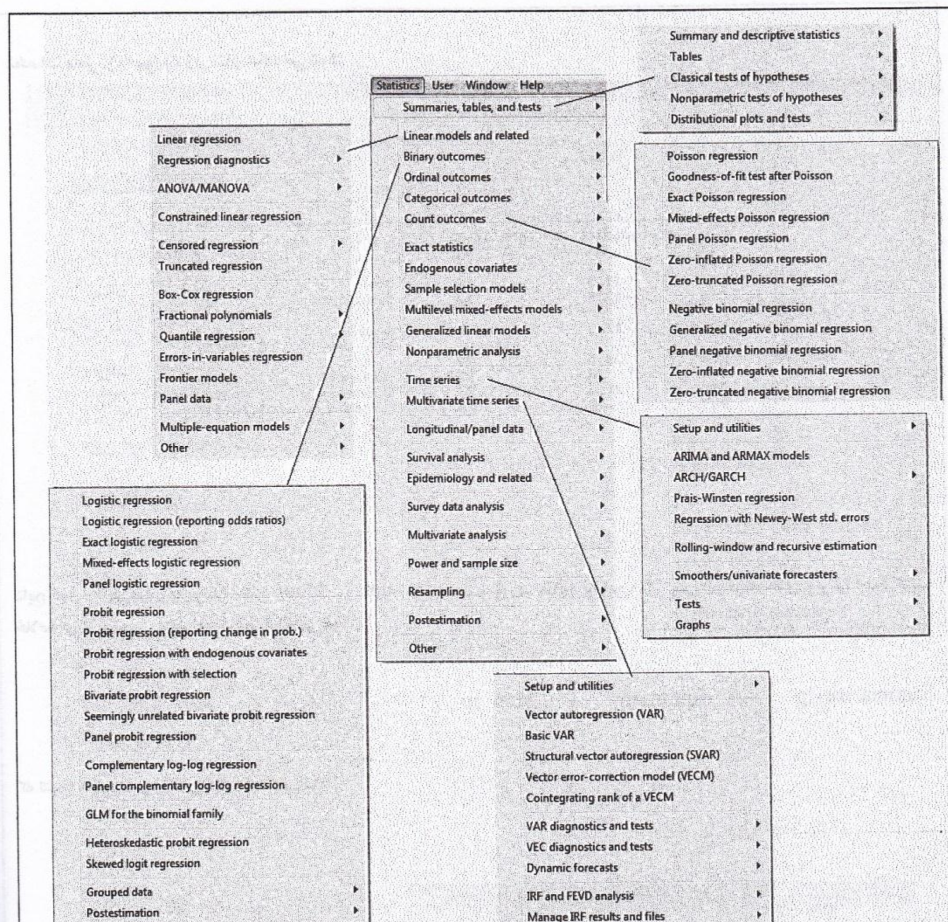
```
sum ip

Variable |      Obs       Mean    Std. Dev.      Min       Max
---------+-----------------------------------------------------
      ip |       21   22847.13    12319.65    4873.2   52803.6
```

نام متغیر، تعداد مشاهدات، میانگین، انحراف معیار، حداقل، حد اکثر

منوی Statistics

محاسبات آماری و روش‌های مربوطه، عمدتاً در گزینه‌های این منو قرار دارد. هر یک از قسمت‌های این منو، جزئیات زیادی دارد که در قسمت‌های بعدی، به آنها اشاره خواهیم کرد. برخی از کارکردهای آن در نمودار زیر نشان داده شده است.



ایجاد متغیر جدید

با استفاده از متغیرهای موجود می‌توان متغیرهای جدیدی را تعریف نمود. بدین منظور از فرمان gen استفاده می‌شود:

```
gen u=y+10
gen v=x+y
gen w=x^2
gen ly=log(y)
gen q=x if x>8
```

آخری، متغیر q را برابر با x قرار می‌دهد ولی به‌ازای مقادیر بزرگتر از هشت. توجه شود که در این حالت، مقدار q در مواردی که x کوچکتر از ۱۰ است، موجود نمی‌باشد. همچنین فرمان‌های دیگری برای سایر محاسبات وجود دارد که به‌عنوان مثال برای محاسبه لگاریتم از فرمان زیر استفاده می‌شود:

```
gen y=log(x)
```

متغیر زمان

متغیر زمان برای داده‌های سالانه: از سال ۱۳۵۰ تا ... با فرمان زیر حساب می‌شود:

```
gen t=1350+_n-1
```

متغیر زمان برای داده‌های فصلی (۳ ماهه): دو فرمان زیر را به‌دنبال هم اجرا می‌کنیم:

```
gen t=tq(1350q1)+_n-1
format t %tq
```

متغیر زمان برای داده‌های ماهانه: دو فرمان زیر را به‌دنبال هم اجرا می‌کنیم:

```
gen t=tm(1350m1)+_n-1
format t %tm
```

متغیر زمان برای داده‌های هفتگی: دو فرمان زیر را به‌دنبال هم اجرا می‌کنیم:

```
gen t=tw(1350w1)+_n-1
format t %tw
```

متغیر زمان برای داده‌های روزانه: دو فرمان زیر را به‌دنبال هم اجرا می‌کنیم:

```
gen t=td(23dec1985)+_n-1
format t %td
```

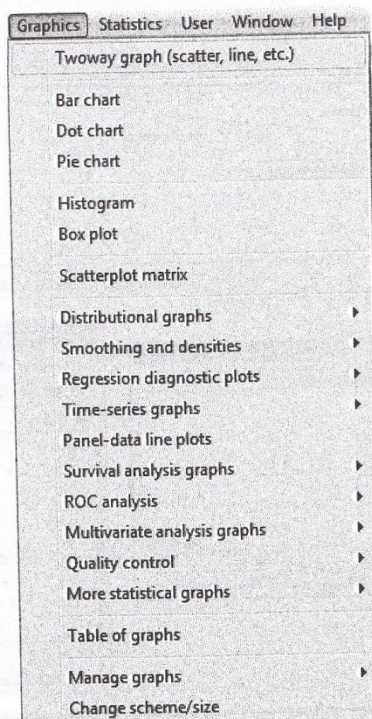
بعد از ایجاد متغیر زمان، فرمان زیر را اجرا می‌کنیم تا متغیر t به‌عنوان زمان در نظر گرفته شود:

```
tsset t
```

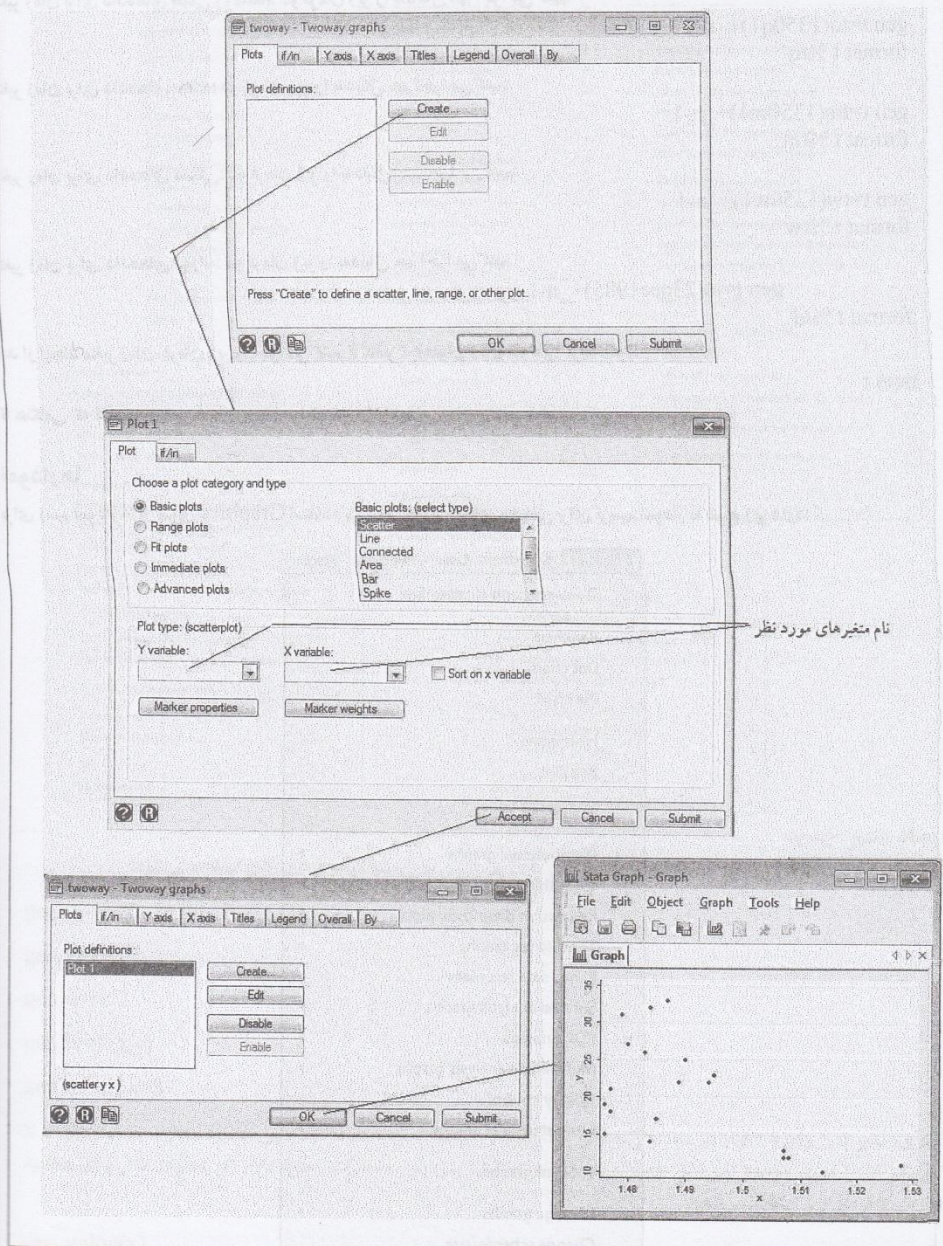
تا هنگامی که هر یک از این فرمان‌ها را اجرا نکنید، stata متغیر زمان را نمی‌شناسد.

نمودارها

برای رسم نمودارها از منوی Graphics استفاده می‌شود که گزینه‌های مختلفی برای ترسیم نمودار به شرح زیر دارد:



اولین گزینه تحت عنوان twoway graph(scater, line, etc) است که بیش از بقیه مورد استفاده می‌باشد. با انتخاب این گزینه، پنجره زیر باز می‌شود:



و یا می‌توان از فرمان‌های کوتاه مانند scatter y x در سطر فرمان استفاده نمود.

همچنین می‌توان نمودار y در مقابل x را بافرمان زیر ترسیم نمود.

```
twoway (scatter y x)
```

برای رسم نمودارهای سری زمانی از فرمان زیر استفاده می‌شود:

```
twoway (line y x t)
```

line نوع نمودار، x و y نام متغیرها، و t مقادیر روی محور x است که می‌تواند زمان یا هر چیز دیگری باشد.

علاوه بر این می‌توان از فرمان زیر برای رسم نمودارهای معمولی مانند سری زمانی استفاده نمود:

```
tsline y x
```

سایر فرمان‌ها

وقفه مرتبه یک (مقدار y در زمان t-1)	l.y یا l1.y
وقفه مرتبه دو (مقدار y در زمان t-۲)	l2.y
مقدار y در زمان t+1	f.y یا f1.y
مقدار y در زمان t+۲	f2.y
تفاصل مرتبه اول (y_t-y_{t-1})	d.y یا d1.y
تفاصل مرتبه اول $(y_t-y_{t-1})-(y_{t-1}-y_{t-2})$	d2.y
تفاصل‌های فصلی (y_t-y_{t-s})	ds.y
(y_t-y_{t-2s})	d2s.y

s=4 برای فصل (سه ماهه) و s=12 برای ماهانه

همچنین برا ی محاسبهٔ مقادیر تأخیری (وقفه) می‌توان از فرمان زیر استفاده نمود:

```
gen y1=y[_n-1]
```

افزودن دامنه

گاهی اوقات برای اضافه کردن داده‌ها و یا برای پیش‌بینی برون‌نمونه‌ای، لازم است که دامنه مورد نظر را گسترش دهیم. بدین منظور می‌توان فرمان زیر را اجرا نمود:

```
tsappend, add(10)
```

فرمان فوق، دامنه موردنظر را ۱۰ تا افزایش می‌دهد که می‌تواند سال، ماه و غیره باشد. مثلاً دامنه را که قبلاً تا سال ۱۳۸۰ بوده تا سال

۱۳۹۰ اضافه می‌کند. علاوه بر این می‌توان از فرمان زیر استفاده نمود.

```
tsappend, last (1380) tsfmt(ty)
```

برای داده‌های فصلی، آخرین فصل (1380q4) را مشخص کرده و به‌جای ty از tq استفاده می‌کنیم. همین فرمان را برای داده‌های ماهانه نیز می‌توان استفاده نمود که در این حالت آخرین زمان مثلاً 1380m12 را معلوم کرده و به‌جای ty از tm استفاده می‌کنیم.

سایر علائم

مساوی ==

مساوی نیست با !=

بزرگتر از >

کوچکتر از <

بزرگتر یا مساوی >=

کوچکتر یا مساوی <=

و &

یا |

لحاظ نکردن برخی از مشاهدات

اگر بخواهیم در تخمین معادلات و یا در سایر محاسبات، شرطی را قائل شویم می‌توان آن را با if لحاظ نمود. به عنوان مثال می‌خواهیم آن بخش از مشاهدات که y برابر صفر است، در نظر گرفته نشوند. بدین منظور از فرمان زیر استفاده می‌کنیم:

```
Drop if y==0
```

این فرمان موجب حذف آن بخش از مشاهدات می‌شود که y=۰ است. همچنین می‌توان به جای صفر هر عدد دیگری را نیز درنظر گرفت. به جای مساوی می‌توان از کوچکتر و بزرگتر نیز استفاده نمود.

حذف یک متغیر

متغیر y را می‌توان با فرمان زیر حذف نمود:

```
drope y
```

همچنین اجرای فرمان زیر، همه متغیرهای یک فایل بجز متغیرهای x، y و z را حذف می‌کند:

```
keep x y z
```

نشان‌گذاری متغیرها

متغیرها که با حروف لاتین نشان داده می‌شوند نیاز به توصیف بیشتری دارند. در این خصوص می‌توان با استفاده از فرمان زیر، یک برچسب را برای متغیر مورد نظر انتخاب نمود. به‌عنوان مثال می‌خواهیم نشان دهیم که y بیانگر رشد GDP است:

```
label  variable y “growth of GDP”
```

با اجرای این فرمان، عبارت growth of GDP مقابل متغیر y ظاهر می‌شود.

فصل دوم

رگرسیون ساده

۲-۱ مقدمه

در تحلیل رگرسیون ساده، تبیین رابطهٔ بین دو متغیر بررسی می‌شود. بر خلاف همبستگی، در اینجا میزان تأثیر یک متغیر بر دیگری اندازه‌گیری می‌شود. به‌عنوان مثال رابطه بین تورم و رشد نقدینگی را می‌توان از طریق محاسبه ضریب همبستگی و یا از طریق رگرسیون بررسی نمود. ضریب همبستگی به بررسی وجود یا عدم وجود رابطه خطی می‌پردازد، اما در تحلیل رگرسیون، یک متغیر (مانند تورم) را تابعی از متغیر دیگر (مانند رشد نقدینگی) در نظر می‌گیریم. بررسی این موارد نیاز به روش‌هایی دارد که اصول آنها در این فصل بررسی می‌شود. بدین منظور در این فصل، ابتدا به مبانی نظری رگرسیون و سپس به روش‌های کاربردی آن خواهیم پرداخت. مبنای تحلیل رگرسیون عمدتاً بر توزیع‌های چند متغیره و توزیع‌های شرطی است که معادله رگرسیون را به‌صورت میانگین شرطی ارائه می‌کند. بدین منظور بحث را با امید ریاضی شرطی شروع می‌کنیم.

۲-۲ امید ریاضی شرطی و رگرسیون

در تحلیل رگرسیون معمولاً یکی از متغیرها را وابسته و دیگری را مستقل فرض می‌کنیم. در ابتدا باید توجه داشت که در تابع $Y = f(X)$ به ازای X، یک مقدار معین برای Y وجود دارد. به‌عبارت دیگر به ازای $X = a$ همواره $Y = f(a)$ خواهد بود. اما در تحلیل رگرسیون به ازای $X = a$، الزاماً همواره $Y = f(a)$ به‌دست نمی‌آید. یعنی اگر X را در سطح a کنترل کنیم،

الزاماً مقدار $Y = f(a)$ به‌دست نمی‌آید، زیرا عوامل دیگری(به‌جز X) وجود دارند که بر Y اثر می‌گذارند. در واقع این بحث بیانگر تفاوت شرایط مطمئن با شرایط نامطمئن می‌باشد. به همین دلیل، وقتی به ازای هر X چندین Y وجود دارد، بایستی از متوسط‌های Y استفاده نمود که معروف به میانگین‌های شرطی یا امید ریاضی شرطی می‌باشد. بنابراین ابتدا لازم است مفهوم امید ریاضی شرطی را بررسی کنیم.

بدین منظور، بحث را با تابع احتمال مشترك پی‌گیری می‌کنیم. فرض کنید تابع احتمال (چگالی) مشترك X و Y به‌صورت زیر باشد:

$$f(x,y) = \frac{۲}{۳}(x+۲y) \qquad ۰<x<۱ \quad , ۰<y<۱$$

توزیع‌های حاشیه‌ای یا توزیع تکی هر یك از متغیرها و همچنین توزیع شرطی Y عبارت است از:

$$f(x) = \int_۰^۱ f(x,y)dy = \int_۰^۱ \frac{۲}{۳} f(x+۲y)dy = \frac{۲}{۳}(۱+x)$$

$$f(y) = \int_۰^۱ f(x,y)dx = \int_۰^۱ \frac{۲}{۳}(x+۲y)dx = \frac{۱}{۳}(۱+۴y)$$

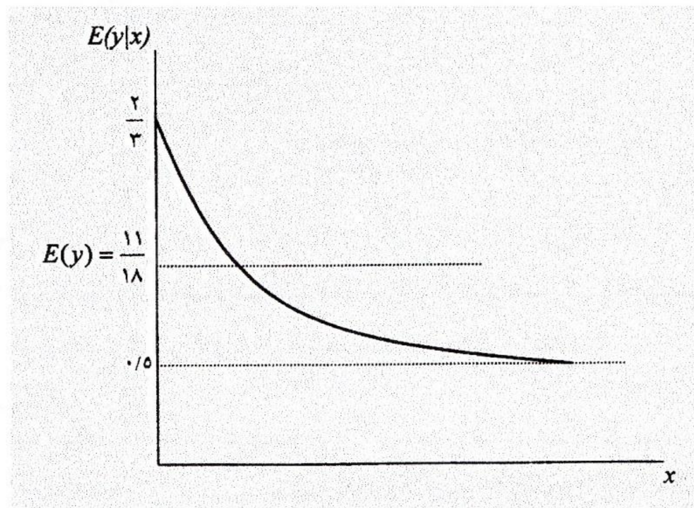
$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{\frac{۲}{۳}(x+۲y)}{\frac{۲}{۳}(x+۱)} = \frac{x+۲y}{x+۱}$$

$f(x)$ و $f(y)$ به ترتیب تابع احتمال X و Y بوده و $f(y|x)$ نیز تابع احتمال شرطی Y می‌باشد. با استفاده از تابع احتمال شرطی می‌توان رابطهٔ بین دو متغیر X و Y را بررسی نمود. اگر شرط $f(y|x)=f(y)$ برقرار باشد، در این صورت این دو متغیر، مستقل‌اند و هیچ رابطه‌ای با هم ندارند. اما در غیر این صورت متغیرهای مذکور، مستقل نیستند. در مثال فوق، چون $f(y|x) \neq f(y)$ است لذا X و Y مستقل نیستند.

اما سؤال این است که چگونه رابطه بین X و Y را به‌صورت کمّی بررسی کنیم. یکی از راههای مناسب این است که امید ریاضی شرطی Y را حساب کنیم. بدیهی است که اگر Y مستقل از X باشد در این صورت $f(y|x)=f(y)$ و $E(Y|X)=E(Y)$ می‌باشد و در نتیجه امید ریاضی شرطی Y تابعی از X نخواهد بود. در مثال فوق امید ریاضی شرطی Y عبارت است از:

$$E(Y|X)=\int_{۰}^{۱} yf(y|x)dy=\int_{۰}^{۱} y\frac{x+۲y}{x+۱}dy=\frac{۳x+۴}{۶(x+۱)}$$

بنابراین همان‌طور که انتظار می‌رفت، امید ریاضی شرطی Y تابعی از X است.



نمودار ۱-۲: امید ریاضی شرطی

پس به‌طورکلی اگر دو متغیر X و Y مستقل نباشند، امید ریاضی شرطی Y تابعی از X خواهد بود:

$$E(Y|X)=g(X) \qquad (۲-۱)$$

معادله (۱-۲) موسوم به معادلهٔ رگرسیون Y روی X می‌باشد. این تابع فقط رابطه بین میانگین‌های شرطی Y با متغیر X بوده که ممکن است خطی یا غیرخطی باشد. اگر خطی باشد، می‌توان آن را در حالت کلی به‌صورت زیر نوشت:

$$E(YX)=\alpha+\beta X \qquad (۲-۲)$$

در کاربردهای تجربی، معمولاً از معادلهٔ خطی استفاده می‌شود و یا اگر معادله‌ای غیرخطی باشد، می‌توان آن را با تقریب، تبدیل به خطی نمود.

رگرسیون نرمال

اگر تابع احتمال مشترک X و Y، نرمال باشد، آنگاه امید ریاضی شرطی Y تابع خطی از X خواهد بود. تابع احتمال مشترک نرمال عبارت است از:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}}$$

تابع احتمال شرطی Y نیز عبارت است از:

$$f(y|x)=\frac{f(x,y)}{f(x)}=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}e^{-\frac{\left[y-\left(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2}{2\sigma_y^2(1-\rho^2)}}$$

X و Y مستقل نیستند، زیرا $f(x,y)\neq f(x)f(y)$ و $f(y|x)\neq f(y)$ می‌باشد.

در اینجا ρ ضریبی است که وابستگی X و Y را بیان می‌کند. اگر $\rho=0$ باشد، آنگاه $f(x,y)=f(x)f(y)$ و $f(y|x)=f(y)$ خواهد شد که بیانگر استقلال X و Y است. امید ریاضی شرطی Y عبارت است از:

$$E(Y|X)=\int_{-\infty}^{+\infty}f(y|x)dx=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(X-\mu_x)$$

$$=(\mu_y-\rho\frac{\sigma_y}{\sigma_x}\mu_x)+\rho\frac{\sigma_y}{\sigma_x}X=\alpha+\beta X$$

$$\alpha=\mu_y-\rho\frac{\sigma_y}{\sigma_x}\quad ,\ \beta=\rho\frac{\sigma_y}{\sigma_x}$$

همچنین واریانس شرطی Y برابر است با:

$$\mathrm{var}(Y|X)=\sigma_{y|x}^2=\int_{-\infty}^{+\infty}\left(y-E(y|x)\right)^2f(y|x)dy=\sigma_y^2(1-\rho^2)$$

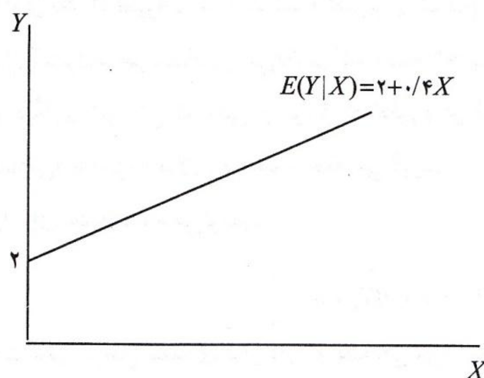
مثال ۱-۲: اگر $\mu_x=5$، $\mu_y=4$، $\sigma_x=4$، $\sigma_y=2$ و $\rho=0/8$ باشد، معادله رگرسیون را رسم کنید.

$$\alpha=4-(0/8\times\frac{2}{4})5=2\ ,\ \beta=0/8\times\frac{2}{4}=0/4$$

$$E(Y|X)=2+0/4X$$

واریانس شرطی Y عبارت است از:

$$\sigma_{y|x}^2=2^2(1-0/8^2)=1/44$$



نمودار ۲-۲: امید ریاضی شرطی خطی در توزیع نرمال

به‌هرحال طبق رابطه (۱-۲) امید ریاضی شرطی Y تابعی از X است. اگر توزیع مشترک X و Y نرمال باشد، آنگاه امید ریاضی شرطی Y تابع خطی از X خواهد بود. اگر امید ریاضی شرطی Y غیرخطی باشد می‌توان آن را با استفاده از بسط توابع با تقریب مرتبه اول، خطی نمود.[1]

به‌طور کلی، امید ریاضی شرطی Y را به‌صورت $E(Y|X) = \alpha + \beta X$ در نظر می‌گیریم. اما بدیهی است که مقادیر واقعی Y با $E(Y \mid X)$ تفاوت دارند و در این خصوص مرتکب خطا می‌شویم. به همین دلیل لازم است که این خطاها را وارد این معادله نماییم.

۲-۳ جملهٔ خطا و معادله رگرسیون

بنا بر آنچه گفته شد معادله (۲-۲) فقط آن بخش از تغییرات Y را توضیح می‌دهد که متأثر از متغیر X می‌باشد. حال تأثیر بقیه عوامل را که تصادفی و غیر قابل کنترل هستند در این معادله وارد می‌کنیم. اثر سایر عوامل را با u نشان می‌دهیم. بنابراین برای هر مقدار از Y می‌توان رابطه زیر را نوشت:

$$Y = E(Y|X) + u \quad \Rightarrow \quad Y = \alpha + \beta X + u \qquad (۲-۳)$$

---

[1] تقریب خطی تابع $g(x)$ در $x = a$ عبارت است از:

$$g(x) \cong g(a) + g'(a)(x-a)$$

که در اینجا $\alpha = g(a) - g'(a)a$ و $\beta = g'(a)$ خواهد بود.

بدین ترتیب طبق الگوی (۳-۲) تغییرات Y به دو دسته تقسیم می‌شود: یکی تغییراتی که متأثر از X است و آن را قابل کنترل و غیرتصادفی می‌دانیم که اصطلاحاً به آن «جزء معین یا غیرتصادفی» می‌گویند و دیگری تغییراتی که ناشی از عوامل تصادفی و غیرقابل کنترل است که اصطلاحاً به آن «جزء تصادفی» یا «جزء اختلال» یا «جمله خطا» می‌گویند.

حال معادله (۳-۲) را برای مشاهده iام می‌نویسیم:

$$Y_i = \alpha + \beta X_i + u_i \qquad (۲-۴)$$

از آنجا که u_i تحت تأثیر عوامل متعددی قرار دارد و تصادفی می‌باشد، لذا u_i یک متغیر تصادفی است. از طرف دیگر هر متغیر تصادفی خصوصیاتی دارد که توسط توزیع آن متغیر و فروض مربوط به آن بیان می‌شود. در ادامه بحث، به بررسی فروض مربوط به جملهٔ اختلال می‌پردازیم.

۲-٤ فروض معادله رگرسیون

فروض معادله رگرسیون یا فروض کلاسیک، خصوصیات جزء تصادفی مدل رگرسیون را تبیین می‌کند. این فروض عبارتند از:

فرض میانگین صفر	۱) $E(u_i\|X)=0$
فرض واریانس همسانی	۲) $\mathrm{var}(u_i\|X)=E(u_i^2\|X)=\sigma^2<\infty$
فرض عدم خودهمبستگی	۳) $\mathrm{cov}(u_k,u_j)=E(u_k u_j)=0 \quad k\neq j$
فرض استقلال جزء تصادفی (u_i) از جزء غیرتصادفی (X_i)	۴) $\mathrm{cov}(u_i,X_i)=0$
فرض نرمال بودن	۵) $u_i \sim \mathrm{N}(0,\sigma^2)$

مفاهیم مربوط به این فروض در مثال زیر ارائه شده است:

فرض کنید که نمرهٔ امتحان دانشجویی (Y) تابعی از میزان ساعات مطالعه وی (X) در روز قبل از امتحان باشد. بدیهی است که عوامل بی‌شمار دیگری نیز دخیل هستند که آنها را با u نشان می‌دهیم، از قبیل این که دوستش به او تلفن بزند و یک ساعت وقت او را بگیرد، مهمان برای آنها بیاید، برق قطع شود، کتابش گم شود و هزاران مسئلهٔ دیگر. اینها عوامل غیرقابل کنترل هستند. همچنین فرض کنید که رابطهٔ $E(Y|X)=4+0/8X$ برقرار باشد. این رابطه بیانگر یک «رابطه

انتظاری» است. مثلاً انتظار بر این است که اگر ۱۰ ساعت مطالعه کند $(X=۱۰)$، نمره وی ۱۲ خواهد شد. اما به‌دلیل وجود عوامل تصادفی ممکن است عملاً به نمرهٔ ۱۲ دست پیدا نکند و به نتیجهٔ دیگری برسد. همان‌طور که گفته شد X متغیر قابل کنترل ولی u متغیر تصادفی است:

$$Y_i = E(Y|X)+u_i = (۴+۰/۸X_i)+u_i \quad \Rightarrow \quad u_i = Y_i-(۴+۰/۸X_i)$$

حال فرض کنید مشاهده kام بدین صورت بوده که او ۱۰ ساعت مطالعه کند و نمرهٔ امتحانش برابر با ۱۲/۵ شده باشد. در این شرایط مشاهده kام به‌صورت $(X_k=۱۰, Y_k=۱۲/۵)$ می‌باشد و آنگاه u_k برابر است با:

$$u_k = Y_k-(۴+۰/۸X_k) = ۱۲/۵-(۴+۰/۸\times۱۰) = ۱۲/۵-۱۲ = ۰/۵$$

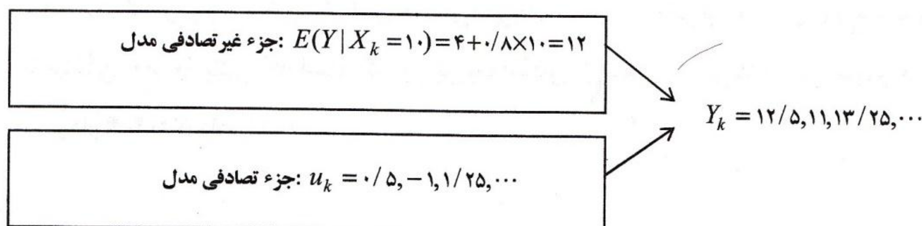
حال تصور کنید که این آزمایش را بتوان تکرار کرده و به روز قبل از امتحان برگردیم و مشاهده kام را یک بار دیگر اندازه‌گیری کنیم. در این حالت قسمت قابل کنترل همچنان مانند قبل است، یعنی $X_k=۱۰$ و $E(Y|X_k=۱۰)=۴+۰/۸\times۱۰=۱۲$ می‌باشد. اما نمره او به‌دلیل وجود عوامل تصادفی برابر با ۱۲/۵ نخواهد شد و به‌عنوان مثال برابر با ۱۱ شود:

$$u_k = Y_k-(۴+۰/۸X_k) = ۱۱-(۴+۰/۸\times۱۰) = ۱۱-۱۲ = -۱$$

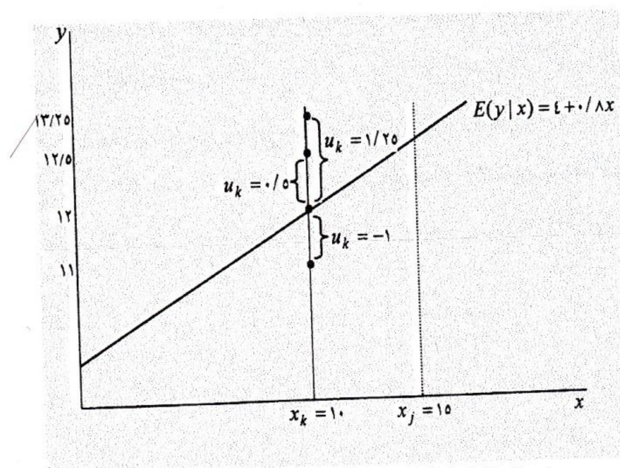
اگر آزمایش را دوباره تکرار کنیم و به روز قبل از امتحان برگردیم و X_k را در سطح ۱۰ کنترل کنیم، این بار نمره او ممکن است به‌عنوان مثال برابر با ۱۳/۲۵ شود:

$$u_k = Y_k-(۴+۰/۸X_k) = ۱۳/۲۵-(۴+۰/۸\times۱۰) = ۱۳/۲۵-۱۲ = ۱/۲۵$$

اگر این آزمایش را مجدداً تکرار کنیم و جزء غیرتصادفی را در سطح $X_k=۱۰$ کنترل کنیم، به Y_kهای متفاوتی می‌رسیم که ناشی از عوامل تصادفی است. در این صورت جزء غیرتصادفی مدل همواره برابر با $E(Y|X_k=۱۰)=۴+۰/۸\times۱۰=۱۲$ خواهد بود. اما جزء تصادفی مدل چون قابل کنترل نمی‌باشد دارای مقادیر غیرقابل پیش‌بینی خواهد بود:



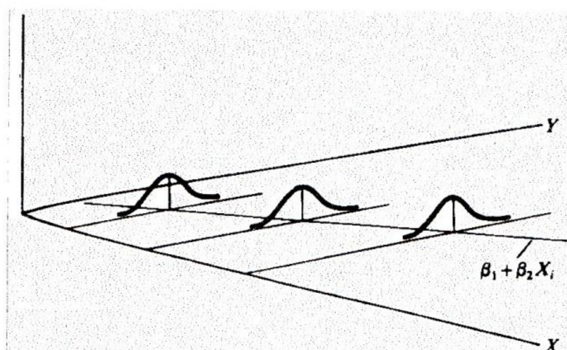
این بحث را روی نمودار نیز می‌توان نشان داد:



نمودار ۳-۲: خطاهای تصادفی و معادله رگرسیون

اگر این آزمایش را به تعداد زیاد تکرار کنیم آنگاه مشاهده خواهد شد که u_k در اطراف عدد ۱۲ (در حالت کلی در اطراف $(\alpha + \beta X_k)$) نوسان می‌کند. در صورت تصادفی بودن u_k، میانگین آن باید صفر شود که یکی از فروض اساسی این مدل است.

اما مفهوم یکسان بودن واریانس u_i این است که وقتی $X_k = ۱۰$ است، پراکندگی (واریانس) u_k برابر با $\sigma^۲$ می‌باشد. حال اگر آزمایش‌های تکراری را در سطح $X_j = ۱۵$ نیز انجام دهیم در این صورت Y_jهای مشاهده شده، ناشی از $X_j = ۱۵$ و u_jها خواهد بود. در اینجا نیز چون $E(Y|X_j = ۱۵) = ۴ + ۰/۸ \times ۱۵ = ۱۶$ است لذا u_jها در اطراف عدد ۱۶ نوسان می‌کند که $E(u_j) = ۰$ و واریانس آن $E(u_j) = \sigma^۲$ خواهد بود. این بدان معنا است که واریانس u_i مستقل از X_i است که موسوم به فرض «واریانس همسانی» می‌باشد. به‌عبارت دقیق‌تر u_i دارای توزیع نرمال با میانگین ۰ و واریانس $\sigma^۲$ است که این توزیع، مستقل از مقدار X_i می‌باشد. این مفهوم در نمـودار ۴-۲ نشان داده شده است:



نمودار ٤-٢

فرض عدم رابطه بین u_iها بدین معنی است که u_k و u_j مستقل‌اند:

$$\operatorname{cov}(u_k, u_j) = E\left[(u_k - Eu_k)(u_j - Eu_j)\right] = E(u_k u_j) = 0 \qquad k \neq j$$

این فرض موسوم به عدم «خودهمبستگی» است که در فصل چهارم بحث خواهد شد. مفهوم این فرض آن است که خطاهای تصادفی، مستقل هستند. به‌عبارت دیگر خطاهای امروز مستقل از خطاهای دیروز است. یا خطاها در یک آزمایش (مشاهده) مستقل از خطاها در آزمایش دیگر است. این بدان معنا است که خطاها باید واقعاً تصادفی باشند و هیچ نظم و ترتیبی در آنها مشاهده نشود. اگر خطاها نظام‌مند (سیستماتیک) باشند آنگاه دچار همبستگی خواهند شد.

فرض غیرتصادفی بودن X بدین معنی است که در آزمایش‌های تکراری، X_k ثابت است و لذا برای آن، $E(X_k) = X_k$ و $\operatorname{var}(X_k) = 0$ خواهد بود[1]. بنابراین بدین نتیجه خواهیم رسید که X_k و u_k مستقل‌اند. به‌عبارت دیگر جزء تصادفی و جزء غیرتصادفی مستقل هستند. حتی اگر X نیز تصادفی باشد، فرض بر این است که فرایندی که داده‌های X را ایجاد می‌کند و فرایندی که u را ایجاد می‌کند، مستقل‌اند.

$$\operatorname{cov}(u_k, X_k) = E[(u_k - Eu_k)(X_k - EX_k)] = E[u_k(X_k - EX_k)] = (Eu_k)E(X_k - EX_k) = 0$$

بر اساس فروض فوق، می‌توان خصوصیات Y_i را نیز بررسی نمود. از آنجا که Y_i تابعی از u_i

---

١- این نکته با این بحث که X پراکندگی ندارد، اشتباه نشود. در اینجا منظور این است که X_i در آزمایش‌های تکراری، ثابت است و لذا $\operatorname{var}(X_i) = 0$ می‌باشد، اما واریانس X که با $\operatorname{var}(X)$ نشان داده می‌شود به صورت زیر می‌باشد:

$$\sigma_x^2 = \operatorname{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

است، لذا Y_i نیز یک متغیر تصادفی است. علاوه بر این، چون u_i نرمال است لذا Y_i که تابع خطی از u_i است، نیز نرمال خوهدبود:

$$۱)\ Y_i=\alpha+\beta X_i+u_i$$

$$۲)\ E(Y|X_i)=\alpha+\beta X_i \quad , \quad E(u_i|X_i)=۰$$

$$۳)\ \text{var}(Y|X_i)=\sigma^۲_{y|x_i}=\text{var}(u_i|X_i)=\sigma^۲$$

$$۴)\ Y_i\sim \text{N}(\alpha+\beta X_i,\sigma^۲)$$

۲-۵ تحلیل رگرسیون تجربی

آنچه تا اینجا بررسی شد مربوط به مبانی نظری تحلیل رگرسیون می‌باشد. اما در مباحث تجربی، صرفاً با مجموعه‌ای از مشاهدات مواجه‌ایم که بر اساس نمونه‌گیری به‌دست آمده است. به‌عنوان مثال در تحلیل رابطهٔ دو متغیر X و Y فقط مجموعه‌ای از داده‌های جفت‌شده را به‌صورت $(X_n,Y_n),\ldots,(X_۲,Y_۲),(X_۱,Y_۱)$ در اختیار داریم.

در تحلیل رگرسیون نظری بدین نتیجه رسیدیم که معادله رگرسیون نظری به‌صورت $E(Y|X_i)=\alpha+\beta X_i$ است. اولاً ممکن است این معادله، خطی نباشد ولی برای سادگی آن را خطی فرض کرده‌ایم که البته در صورت برقراری فرض نرمال بودن u_i، به تبع آن Y_i نیز توزیع نرمال خواهد داشت و فرض خطی بودن می‌تواند واقعی باشد. ثانیاً معادله رگرسیون $E(Y|X_i)=\alpha+\beta X_i$ بیانگر مقدار انتظاری Y_i است نه مقدار دقیق آن. در واقع برآورد ما از Y_i طبق $\alpha+\beta X_i$ صورت می‌گیرد که برای ما قابل کنترل است. در حالی که مقدار واقعی Y_i برابر با $Y_i=\alpha+\beta X_i+u_i$ است. ثالثاً معادله رگرسیون $E(Y|X_i)=\alpha+\beta X_i$ را نمی‌دانیم، زیرا α و β مجهول هستند. بنابراین مشکل اصلی این است که چگونه α و β را تعیین کنیم. فرض کنید که به هر طریقی α و β را برآورد کرده و آن را با $\hat{\alpha}$ و $\hat{\beta}$ نشان دهیم. در این صورت برآورد ما از Y_i برابر با $\hat{\alpha}+\hat{\beta}X_i$ خواهد بود که متفاوت از $\alpha+\beta X_i$ است. بنابراین اگر برآوردی را که طبق معادله $\hat{\alpha}+\hat{\beta}X_i$ به‌دست می‌آوریم با $\hat{Y}_i$ نشان دهیم آنگاه در رسیدن به Y_i دچار خطایی به اندازه $Y_i-\hat{Y}_i$ خواهیم شد که آن را با e_i نشان می‌دهیم. لذا در تحلیل رگرسیون تجربی، رابطه $\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i$ و $Y_i=(\hat{\alpha}+\hat{\beta}X_i)+e_i$ را داریم. در واقع به‌جای $\alpha+\beta X_i$ از $\hat{\alpha}+\hat{\beta}X_i$ و

به‌جای u_i از e_i استفاده می‌کنیم. به‌طور خلاصه، تحلیل رگرسیون نظری و تجربی در جدول زیر نشان داده شده است:

مقدار خطا		مقدار تخمینی		مقدار واقعی
خطای تخمینی	خطای واقعی	رگرسیون تجربی	رگرسیون نظری	
$\hat{u}_i = e_i = Y_i - \hat{Y}_i$ $= Y_i - (\hat{\alpha} + \hat{\beta}X_i)$	$u_i = Y_i - E(Y \mid X_i)$ $= Y_i - (\alpha + \beta X_i)$	$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$	$E(Y \mid X_i) = \alpha + \beta X_i$	Y_i

به‌هرحال چون α و β را نمی‌دانیم لذا امکان دسترسی به معادلهٔ $E(Y \mid X_i) = \alpha + \beta X_i$ وجود ندارد و به ناچار باید از تخمین α و β استفاده کنیم تا بدین طریق معادله $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ را جایگزین معادلهٔ $E(Y \mid X_i) = \alpha + \beta X_i$ کنیم. بدین ترتیب بحث بعدی ما راجع به برآورد ضرایب α و β خواهد بود.

۶-۲ تخمین معادله رگرسیون

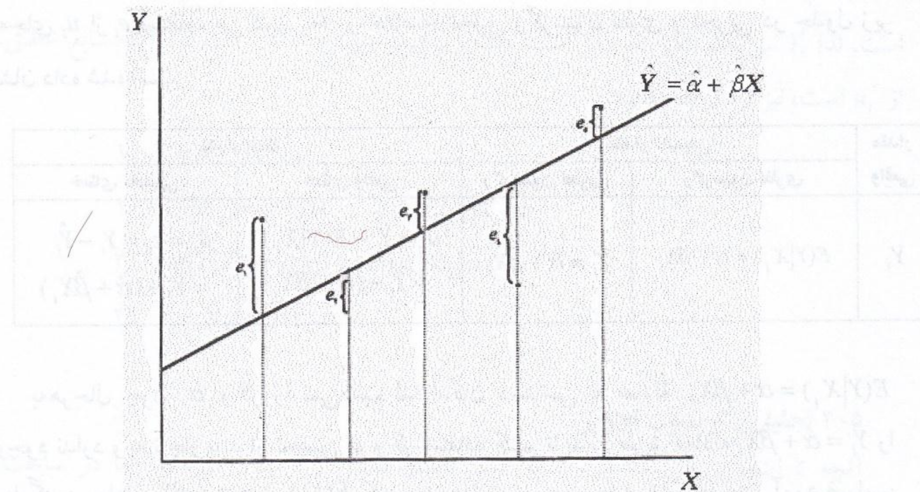
همان‌طور که در بخش قبلی اشاره شد، در مطالعات کاربردی صرفاً داده‌های نمونه را داریم که شامل $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ می‌باشد. حال با استفاده از این داده‌ها، ضرایب α و β را برآورد می‌کنیم. فرض کنید که $\hat{\alpha}$ و $\hat{\beta}$ تخمین α و β باشند. در این صورت روابط زیر را داریم:

$$\text{مقدار واقعی} = Y_i$$

$$\text{مقدار تخمینی} = \hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \qquad (۲-۵)$$

$$\text{مقدار خطا} = \hat{u}_i = e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta}X_i) \qquad (۲-۶)$$

یکی از معیارهای مناسب برای تعیین $\hat{\alpha}$ و $\hat{\beta}$ این است که خطا را حداقل کنیم. بدین منظور اگر مشاهدات X و Y را روی نمودار ترسیم کنیم، نقاط پراکنده‌ای به‌دست می‌آید که باید خط رگرسیون به‌گونه‌ای از بین این نقاط بگذرد که کمترین خطا را به‌وجود آورد. فرض کنید که ۵ مشاهده برای X و Y داریم که در نمودار ۵-۲ ترسیم شده است.



نمودار ۵-۲: خط رگرسیون تجربی

از آنجا که خطاهای مثبت و منفی همدیگر را خنثی می‌کنند، لذا مجموع خطاها، یعنی $\sum e_i$ برابر صفر است. بدین منظور از مجموع مجذور خطاها استفاده می‌شود:

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \ldots + e_n^2 \qquad (2-7)$$

در تخمین $\hat{\alpha}$ و $\hat{\beta}$ بایستی مجموع مربعات خطا، حداقل شود. این روش، معروف به روش حداقل مربعات معمولی (OLS) [1] است. با جایگذاری از (۶-۲) در (۷-۲) خواهیم داشت:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 \qquad (2-8)$$

عبارت فوق وقتی حداقل می‌شود که خط رگرسیون به‌گونه‌ای ترسیم شود که در مجموع، کمترین خطا را ایجاد کند. بدیهی است که موقعیت خط رگرسیون بستگی به عرض از مبدأ آن ($\hat{\alpha}$) و شیب آن ($\hat{\beta}$)دارد. لذا $\hat{\alpha}$ و $\hat{\beta}$ را به‌گونه‌ای تعیین می‌کنیم که مجموع مجذور خطا را حداقل نماید. این کار را می‌توان با مشتق‌گیری نسبت‌به $\hat{\alpha}$ و $\hat{\beta}$ انجام داد:

[1] Ordinary Least Squares

$$\frac{\partial \sum e_i^2}{\partial \hat{\alpha}} = -2\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}} = -2\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)(-X_i) = 0$$

با ساده نمودن این معادلات، خواهیم داشت:

$$\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \quad \text{یا} \quad \sum e_i = 0 \qquad (۲-۹)$$
$$\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i = 0 \quad \text{یا} \quad \sum e_i X_i = 0$$

معادلات (۹-۲) معروف به معادلات نرمال هستند. با حل معادله اول خواهیم داشت:

$$\sum Y_i - n\hat{\alpha} - \hat{\beta}\sum X_i = 0 \Rightarrow \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

حال به‌جای $\hat{\alpha}$ در معادله دوم قرار داده و $\hat{\beta}$ را به‌دست می‌آوریم:[1]

$$\sum[Y_i - (\overline{Y} - \hat{\beta}\overline{X}) - \hat{\beta}X_i]X_i = 0 \Rightarrow \sum[(Y_i - \overline{Y}) - \hat{\beta}(X_i - \overline{X})]X_i = 0$$
$$\Rightarrow \sum[(Y_i - \overline{Y})X_i - \hat{\beta}(X_i - \overline{X})X_i] = 0$$

با حل معادله فوق برای $\hat{\beta}$ خواهیم داشت:

$$\hat{\beta} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} = \frac{\sum X_i Y_i - n\overline{XY}}{\sum X_i^2 - n\overline{X}^2} \qquad (۲-۱۰)$$

از اینجا به بعد، حروف بزرگ را برای مقادیر معمولی متغیرها و حروف کوچک را برای انحراف از میانگین، استفاده می‌کنیم. لذا $X_i - \overline{X} = x_i$ و $Y_i - \overline{Y} = y_i$ می‌باشد. بر این اساس، $\hat{\beta}$ برابر است با:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

که $\sum x_i^2 = \sum(X_i - \overline{X})^2 = \sum X_i^2 - n\overline{X}^2$ و $\sum x_i y_i = \sum(X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - n\overline{XY}$ است.

---

[1] توجه شود که روابط زیر برقرار است:

$$\sum(X_i - \overline{X})X_i = \sum(X_i - \overline{X})^2$$
$$\sum(Y_i - \overline{Y})X_i = \sum(X_i - \overline{X})(Y_i - \overline{Y})$$

مثال۲-۲: مشاهدات X شامل ۱، ۲، ۳، ۳، ۲، ۴، ۶، ۷، ۸، ۶، ۹، ۸، ۷، ۸، ۹، ۹، ۱۰، ۱۰، ۷ و ۹ و Y شامل ۵، ۶، ۵، ۷، ۸، ۹، ۱۲، ۱۰، ۱۰، ۱۲، ۱۴، ۱۵، ۱۵، ۱۰، ۱۴، ۱۶، ۱۸، ۲۰، ۱۸ و ۱۸ می‌باشد. معادله رگرسیون Y روی X را برآورد کنید.

X_i	Y_i	$X_i^۲$	$Y_i^۲$	X_iY_i
۱	۵	۱	۲۵	۵
۲	۶	۴	۳۶	۱۲
۳	۵	۹	۲۵	۱۵
۳	۷	۹	۴۹	۲۱
۲	۸	۴	۶۴	۱۶
۴	۹	۱۶	۸۱	۳۶
۶	۱۲	۳۶	۱۴۴	۷۲
۷	۱۰	۴۹	۱۰۰	۷۰
۸	۱۰	۶۴	۱۰۰	۸۰
۶	۱۲	۳۶	۱۴۴	۷۲
۹	۱۴	۸۱	۱۹۶	۱۲۶
۸	۱۵	۶۴	۲۲۵	۱۲۰
۷	۱۵	۴۹	۲۲۵	۱۰۵
۸	۱۰	۶۴	۱۰۰	۸۰
۹	۱۴	۸۱	۱۹۶	۱۲۶
۹	۱۶	۸۱	۲۵۶	۱۴۴
۱۰	۱۸	۱۰۰	۳۲۴	۱۸۰
۱۰	۲۰	۱۰۰	۴۰۰	۲۰۰
۷	۱۸	۴۹	۳۲۴	۱۲۶
۹	۱۸	۸۱	۳۲۴	۱۶۲
۱۲۸	۲۴۲	۹۷۸	۳۳۳۸	۱۷۶۸

$$\overline{X} = \frac{۱۲۸}{۲۰} = ۶/۴ \quad , \quad \overline{Y} = \frac{۲۴۲}{۲۰} = ۱۲/۱$$

$$\sum x_i^۲ = ۹۷۸ - ۲۰(۶/۴)^۲ = ۱۵۸/۸$$

$$\sum y_i^۲ = ۳۳۳۸ - ۲۰(۱۲/۱)^۲ = ۴۰۹/۸$$

$$\sum x_i y_i = ۱۷۶۸ - ۲۰(۶/۴)(۱۲/۱) = ۲۱۹/۲$$

$$\hat{\beta} = \frac{۲۱۹/۲}{۱۵۸/۸} = ۱/۳۸ \quad , \quad \hat{\alpha} = ۱۲/۱ - ۱/۳۸ \times ۶/۴ = ۳/۲۶۶$$

تخمین معادله رگرسیون عبارت است از:

$$\hat{Y}_i = ۳/۲۶۶ + ۱/۳۸ X_i$$

۷-۲ خواص تخمین‌زننده‌های OLS

هر تخمین‌زننده‌ای بایستی برخی خواص مطلوب را داشته باشد. در این راستا، تخمین‌زننده‌های OLS را اصطلاحاً «بهترین تخمین خطی بدون تورش»[1] می‌گویند. در اینجا هر یک از این خواص را بررسی می‌کنیم.

[1]Best Linear Unbaised Estimation (BLUE)

۲-۷-۱ تخمین‌زنندهٔ خطی

$\hat{\beta}$ را می‌توان به‌صورت زیر نوشت[1]:

$$\hat{\beta}=\frac{\sum x_i y_i}{\sum x_i^{۲}}=\frac{\sum x_i Y_i}{\sum x_i^{۲}}=\sum w_i Y_i$$
$$=\sum w_i(\alpha+\beta X_i+u_i)=\beta+\sum w_i u_i \quad ; \quad w_i=\frac{x_i}{\sum x_i^{۲}} \qquad (۲-۱۱)$$

در اینجا از روابط $\sum w_i = ۰$ و $\sum w_i X_i = ۱$ استفاده شده است. بدین‌ترتیب ملاحظه می‌شود که $\hat{\beta}$ تابع خطی از متغیرهای تصادفی نرمال است و لذا $\hat{\beta}$ نیز یک متغیر تصادفی نرمال می‌باشد.

۲-۷-۲ بدون تورش بودن

با توجه به فرض غیر تصادفی بودن X_i و استقلال آن از u_i، خواهیم داشت:

$$E(\hat{\beta})=E(\beta+\sum w_i u_i)=\beta+\sum w_i E(u_i)=\beta \qquad (۲-۱۲)$$

زیرا $E(u_i)=۰$ است.

۲-۷-۳ سازگاری

سازگاری یک تخمین‌زنندهٔ بدون تورش بیانگر آن است که با افزایش حجم نمونه، واریانس آن به‌سمت صفر میل کند. بدین منظور واریانس $\hat{\beta}$ را حساب می‌کنیم. با استفاده از (۲-۱۱) داریم:

$$\mathrm{var}(\hat{\beta})=E(\hat{\beta}-\beta)^{۲}=E(\sum w_i u_i)^{۲}=\sum_{i=۱}^{n}E(w_i u_i)^{۲}+\sum_{\substack{i=۱\\ i\neq j}}^{n}\sum_{j=i}^{n}E(w_i w_j u_i u_j)$$
$$=\sum_{i=۱}^{n}w_i^{۲}E(u_i^{۲})+\sum_{\substack{i\\ i\neq j}}\sum_{j}w_i w_j E(u_i u_j)=\sum_{i=۱}^{n}w_i^{۲}\sigma^{۲}=\sigma^{۲}\sum_i w_i^{۲}=\frac{\sigma^{۲}}{\sum x_i^{۲}} \qquad (۲-۱۳)$$

زیرا $E(u_i u_j)$ طبق فرض عدم خودهمبستگی u_iها، برابر با صفر است[2]. واریانس $\hat{\beta}$ را به‌صورت زیر بازنویسی می‌کنیم:

---

[1] روابط $\sum x_i=۰$ ، $\sum x_i X_i=\sum x_i^{۲}$ و $\sum x_i y_i=\sum x_i Y_i=\sum X_i y_i$ می‌باشد.

[2] توجه شود که $\sum w_i^{۲}$ برابر است با:

$$\sum w_i^{۲}=\sum\left(\frac{x_i}{\sum x_i^{۲}}\right)^{۲}=\frac{\sum x_i^{۲}}{\left(\sum x_i^{۲}\right)^{۲}}=\frac{۱}{\sum x_i^{۲}}$$

$$\operatorname{var}(\hat{\beta})=\frac{\sigma^{2}/n}{\sum x_{i}^{2}/n}=\frac{\sigma^{2}/n}{S_{x}^{2}}$$

چون S_x^2 مقدار معینی دارد و با افزایش حجم نمونه نیز مقدار آن معلوم و معین است، لذا خواهیم داشت:

$$\lim_{n\to\infty}\operatorname{var}(\hat{\beta})=\frac{\lim\sigma^{2}/n}{\lim S_{x}^{2}}=0$$

زیرا $\lim\sigma^2/n$ برابر صفر است. بنابراین، $\hat{\beta}$ یک تخمین‌زنندهٔ سازگار است.

۲-۷-۴ حداقل واریانس

قضیه گوس- مارکف بیان می‌دارد که از بین تمامی تخمین‌زننده‌های خطی و بدون تورش، روش OLS کمترین واریانس را دارد. بدین منظور، مشابه با (۲-۱۱) تخمین‌زنندهٔ خطی دیگری مانند $\tilde{\beta}$ را در نظر بگیرید:

$$\tilde{\beta}=\sum a_{i}Y_{i}=\sum a_{i}(\alpha+\beta X_{i}+u_{i})=\alpha\sum a_{i}+\beta\sum a_{i}X_{i}+\sum a_{i}u_{i}$$

امید ریاضی $\tilde{\beta}$ عبارت است از:

$$E(\tilde{\beta})=\alpha\sum a_{i}+\beta\sum a_{i}X_{i}$$

برای اینکه $\tilde{\beta}$ بدون تورش باشد بایستی شرط $\sum a_i=0$ و $\sum a_iX_i=1$ برقرار باشد.

واریانس $\tilde{\beta}$ برابر است با:

$$\operatorname{var}(\tilde{\beta})=E(\tilde{\beta}-\beta)^{2}=E\left(\sum a_{i}u_{i}\right)^{2}=E\left(\sum a_{i}^{2}u_{i}^{2}\right)=\sigma^{2}\sum a_{i}^{2}$$

با توجه به اینکه a_i با w_i متفاوت است، لذا آن را به‌صورت $a_i=w_i+b_i$ می‌نویسیم (توجه شود که چون $\sum a_i=\sum w_i=0$ و $\sum a_iX_i=\sum w_iX_i=1$ است، لذا بایستی $\sum b_iX_i=0$ باشد):

$$\begin{aligned}\operatorname{var}(\tilde{\beta})&=\sigma^{2}\sum a_{i}^{2}=\sigma^{2}\sum(w_{i}+b_{i})^{2}\\&=\sigma^{2}\sum(w_{i}^{2}+b_{i}^{2})=\sigma^{2}\sum w_{i}^{2}+\sigma^{2}\sum b_{i}^{2}=\operatorname{var}(\hat{\beta})+\sigma^{2}\sum b_{i}^{2}\end{aligned}$$

چون $\sigma^2\sum b_i^2$ غیر منفی است لذا $\operatorname{var}(\tilde{\beta})\geq\operatorname{var}(\hat{\beta})$ می‌باشد و نتیجه می‌شود که تخمین‌زنندهٔ OLS کاراتر است.

۲-۷-۵ توزیع تخمین‌زننده‌های OLS

امید ریاضی و واریانس $\hat{\beta}$ نشان می‌دهد که تخمین‌زنندهٔ بدون تورش و با حداقل واریانس برای β است. از طرف دیگر چون $\hat{\beta}$ تابع خطی از u_i است لذا $\hat{\beta}$ نیز یک متغیر تصادفی با توزیع نرمال می‌باشد. مشابه این محاسبات را برای $\hat{\alpha}$ نیز می‌توان انجام داد[1]. بدین ترتیب مشخصات تخمین‌زننده‌های OLS عبارت است از:

$$\hat{\beta}=\beta+\sum w_i u_i \quad , \quad w_i=\sum\left(\frac{x_i}{\sum x_i^{۲}}\right)$$

$$\hat{\alpha}=\alpha+\sum k_i u_i \quad , \quad k_i=\frac{۱}{n}-\overline{X}\frac{x_i}{\sum x_i^{۲}} \qquad (۲-۱۴)$$

$$\hat{\beta}\sim \mathrm{N}\left(\beta,\frac{\sigma^{۲}}{\sum x_i^{۲}}\right) \quad , \quad \hat{\alpha}\sim \mathrm{N}\left(\alpha,(\frac{۱}{n}+\frac{\overline{X}}{\sum x_i^{۲}})\sigma^{۲}\right)$$

۲-۷-٦ همبستگی $\hat{\alpha}$ و $\hat{\beta}$

همبستگی تخمین‌زننده‌های $\hat{\alpha}$ و $\hat{\beta}$ باعث می‌شود که تغییر یکی بر دیگری تأثیر بگذارد. بدین منظور کوواریانس $\hat{\alpha}$ و $\hat{\beta}$ را حساب می‌کنیم:

$$\mathrm{cov}(\hat{\alpha},\hat{\beta})=E\left[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\right]$$

به‌جای $\hat{\alpha}-\alpha$ و $\hat{\beta}-\beta$ از (۲-۱۴) قرار می‌دهیم:

$$E\left[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\right]=E\left[\sum_i k_i u_i\right]\left[\sum_i w_i u_i\right]$$

چون امیدریاضی $u_i u_j$ به‌ازای $i\neq j$ برابر صفر است، لذا خواهیم داشت:

$$E\left[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\right]=E\left[\sum_i k_i w_i u_i^{۲}\right]=\sigma^{۲}\sum_i k_i w_i u_i$$

به‌جای k_i و w_i از (۴-۱۴) قرار داده و آن را ساده می‌کنیم که طبق آن، خواهیم داشت:

---

[1] برای محاسبهٔ امید ریاضی و واریانس $\hat{\alpha}$ لازم است آن را به صورت زیر بنویسیم:

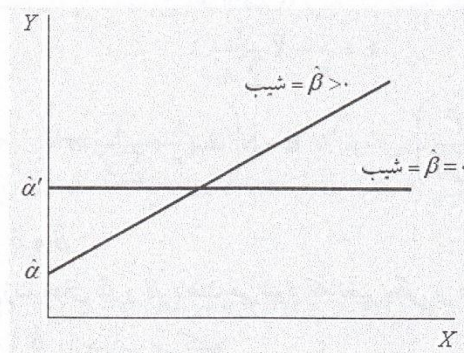
$$\hat{\alpha}=\overline{Y}-\overline{X}\hat{\beta}=\frac{\sum Y_i}{n}-\overline{X}\frac{\sum x_i Y_i}{\sum x_i^{۲}}$$

$$=\sum\left(\frac{۱}{n}-\overline{X}\frac{x_i}{\sum x_i^{۲}}\right)Y_i=\sum k_i Y_i \ , \ k_i=\frac{۱}{n}-\frac{\overline{X}x_i}{\sum x_i^{۲}}$$

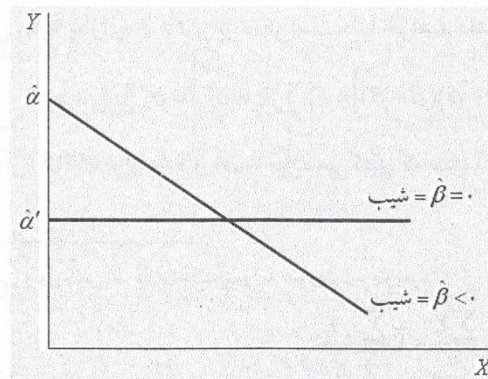
$$=\sum k_i(\alpha+\beta X_i+u_i)=\alpha+\sum k_i u_i \ , \ \sum k_i X_i=۰ \ , \ \sum k_i=۱$$

$$\text{cov}(\hat{\alpha},\hat{\beta}) = -\overline{X}\frac{\sigma^2}{\sum x_i^2} = -\overline{X}\,\text{var}(\hat{\beta}) \qquad (۲-۱۵)$$

مقدار این کوواریانس، منفی است و بدان معنا است که مثلاً اگر X را حذف کنیم معادل با آن است که β را برابر با صفر قرار داده‌ایم و این منجر به افزایش $\hat{\alpha}$ می‌شود (البته مشروط به آن که $\beta \geq 0$ باشد). یعنی یک معادله با شیب مثبت را تبدیل به یک خط افقی با شیب صفر کرده‌ایم که عرض از مبدأ آن افزایش می‌یابد.



ولی اگر $\beta \leq 0$ باشد،آنگاه حذف X و صفر قرار دادن β موجب می‌شود که یک رگرسیون نزولی را تبدیل به یک رگرسیون افقی کنیم. بنابراین، β را افزایش داده‌ایم (از منفی به صفر) و لذا $\hat{\alpha}$ کاهش می‌یابد (چون کوواریانس $\hat{\alpha}$ و $\hat{\beta}$ منفی است).



همچنین می‌توان اثر حذف $\hat{\alpha}$ بر $\hat{\beta}$ را نیز بررسی نمود که این موضوع در فصل چهارم بررسی شده است.

۸-۲ تغییرات کل[1]، تغییرات توضیح داده شده[2] و تغییرات توضیح داده نشده[3]

تغییرات کل Y بیانگر مجموع مجذور تغییرات Y حول $\bar{Y}$ است:

$$\text{تغییرات کل } Y = \sum (Y_i - \bar{Y})^2 = \sum y_i^2 = TSS$$

تغییرات توضیح داده شده بیانگر آن بخش از تغییرات Y است که توسط معادله رگرسیون، توضیح داده می‌شود:

$$\text{تغییرات توضیح داده شدهٔ } Y \text{ توسط معادله رگرسیون} = \sum (\hat{Y}_i - \bar{Y})^2 = \sum \hat{y}_i^2 = ESS$$

تغییرات توضیح داده نشده بیانگر آن بخش از تغییرات Y است که ناشی از سایر عوامل است که برابر با مجموع مجذور خطاها است:

$$\text{تغییرات توضیح داده نشده یا خطاها یا باقیمانده‌ها} = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2 = RSS$$

بدیهی است که رابطه $TSS = ESS + RSS$ برقرار است. بدین منظور توجه شود که اگر از $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ به‌جای $\hat{\alpha}$ در معادله $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ قرار دهیم، رابطه $\hat{Y}_i - \bar{Y} = \hat{\beta}(X_i - \bar{X})$ به‌دست می‌آید که بر اساس آن، ESS عبارت است از:

$$ESS = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}^2 \sum x_i^2 = \hat{\beta} \sum x_i y_i$$

از طرف دیگر اگر در معادله $Y_i = \hat{Y}_i + e_i = \hat{\alpha} + \hat{\beta}X_i + e_i$ به‌جای $\hat{\alpha}$ قرار دهیم، نتیجهٔ $Y_i - \bar{Y} = \hat{\beta}(X_i - \bar{X}) + e_i$ یا $y_i = \hat{\beta}x_i + e_i$ به‌دست می‌آید. لذا TSS عبارت است از:

$$\begin{aligned} TSS &= \sum (Y_i - \bar{Y})^2 = \sum y_i^2 = \sum (\hat{\beta} x_i + e_i)^2 \\ &= \hat{\beta}^2 \sum x_i^2 + \sum e_i^2 + 2\hat{\beta} \sum x_i e_i \\ &= \hat{\beta}^2 \sum x_i^2 + \sum e_i^2 = ESS + RSS \end{aligned} \qquad (۲-۱۶)$$

عبارت $\sum x_i e_i$ برابر با صفر است، زیرا $\sum x_i e_i = \sum (X_i - \bar{X}) e_i = \sum e_i X_i - \bar{X} \sum e_i$ بوده و طبق معادلات (۹-۲)، این دو جمله برابر با صفر می‌باشند.

---

[1] TSS=Total Sum of Squares
[2] ESS= Explaind Sum of Squares
[3] RSS=Residual Sum of Squares

۲-۹ ضریب تعیین ($R^۲$)

بعد از تخمین معادله رگرسیون، اولین سؤال این است که تا چه اندازه تخمین $\hat{Y}_i$ به Y_i نزدیک است. به‌عبارت دیگر، چقدر معادله رگرسیون تخمینی، معادله خوبی است و می‌تواند تغییرات Y را توضیح دهد. در اینجا از معیاری به نام «ضریب تعیین» استفاده می‌شود که بیانگر نسبت تغییرات توضیح داده شده (ESS) به تغییرات کل است:

$$R^۲ = \frac{ESS}{TSS}$$

با جایگذاری به‌جای TSS و TSS خواهیم داشت:

$$R^۲ = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^۲}{\sum y_i^۲} = \frac{\hat{\beta}^۲ \sum x_i^۲}{\sum y_i^۲} = \frac{\hat{\beta} \sum x_i y_i}{\sum y_i^۲} \qquad (۲-۱۷)$$

اگر به‌جای $\hat{\beta}^۲$ قرار دهیم نتیجه می‌شود که ضریب تعیین برابر با مجذور ضریب همبستگی است. همچنین می‌توان ضریب تعیین را به‌صورت زیر نوشت:

$$R^۲ = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = ۱ - \frac{RSS}{TSS} = ۱ - \frac{\sum e_i^۲}{\sum y_i^۲}$$

بدین ترتیب $R^۲$ بیانگر قدرت توضیح دهندگی یا ضریب خوبیِ مدل می‌باشد.

همان‌طور که گفته شد، $R^۲$ معیاری برای قدرت توضیح‌دهندگی معادله رگرسیون است. یعنی نشان می‌دهد که $\hat{Y}_i$ها تا چه اندازه همسو با Y_iها هستند. لذا $R^۲$ برابر با مجذور ضریب همبستگی بین $\hat{Y}_i$ و Y_i است. بدین منظور ابتدا ESS را به‌صورت زیر می‌نویسیم:

$$ESS = \sum \hat{y}_i^۲ = \sum \hat{y}_i \hat{y}_i = \sum \hat{y}_i (y_i - e_i) \quad ; \quad y_i = \hat{y}_i + e_i$$

$$= \sum \hat{y}_i y_i + \sum \hat{y}_i e_i = \sum \hat{y}_i y_i$$

که $\sum \hat{y}_i e_i = ۰$ است.[1] حال $R^۲$ را به‌صورت زیر می‌نویسیم:

---

[1] با استفاده از $\hat{y}_i = \beta x_i$ خواهیم داشت:

$$\sum \hat{y}_i e_i = \sum \hat{\beta} x_i e_i = \hat{\beta} \sum e_i x_i = ۰$$

طبق معادلات نرمال، $\sum e_i x_i = ۰$ است.

$$R^{۲}=\frac{ESS}{TSS}=\frac{(ESS)^{۲}}{(TSS)(ESS)}=\frac{(\sum \hat{y}_i y_i)^{۲}}{\sum (y_i^{۲})(\sum \hat{y}_i^{۲})}=r_{\hat{y}y}^{۲}$$

بنابراین $R^{۲}$ برابر با مجذور ضریب همبستگی بین $\hat{Y}$ و Y است.

مثال۳-۲: در مثال ۲-۲ ضریب تعیین را محاسبه کنید:

$$R^{۲}=\frac{ESS}{TSS}$$

$$ESS=\hat{\beta}^{۲}\sum x_i^{۲}=(۱/۳۶)^{۲}(۱۵۸/۸)=۳۰۲/۵۷$$

$$TSS=\sum y_i^{۲}=۳۳۳۸-۲۰(۱۲/۱)^{۲}=۴۰۹/۸$$

$$R^{۲}=\frac{۳۰۲/۵۷}{۴۰۹/۸}=۰/۷۳۸$$

بنابراین رگرسیون تخمینی توانسته است حدود ۷۳/۸ درصد از تغییرات Y را توضیح دهد و بقیه آن ناشی از عوامل تصادفی (u) می‌باشد.

۲-۱۰ میانگین خطای تخمین یا انحراف معیار رگرسیون($\hat{\sigma}$)

میانگین خطای تخمین یا انحراف معیار رگرسیون، میزان پراکندگی مشاهدات را در اطراف خط رگرسیون نشان می‌دهد. وقتی خط رگرسیون به مشاهدات واقعی نزدیکتر باشد، خطای کمتری دارد و لذا انحراف معیار کوچکتر است. به‌طورکلی، انحراف معیار معادلات رگرسیون بیانگر متوسط خطای معادله رگرسیون است. خطای معادله رگرسیون را می‌توان با محاسبه واریانس شرطی Y یا واریانس جملهٔ خطا، محاسبه نمود که برابر با تخمین $\hat{\sigma}^{۲}$ است.

برای محاسبه $\hat{\sigma}_{۲}$ لازم است که ابتدا $\sum(e_i-\bar{e})^{۲}=\sum e_i^{۲}$ را حساب کنیم. بدین منظور معادله $Y_i=\alpha+\beta X_i+u_i$ را جمع زده و به‌صورت $\sum Y_i=n\alpha+\beta\sum X_i+\sum u_i$ می‌نویسیم که با تقسیم آن بر n، معادله $\bar{Y}=\alpha+\beta\bar{X}+\bar{u}$ به‌دست می‌آید. حال معادله زیر را به‌دست می‌آوریم:

$$Y_i-\bar{Y}=(\alpha+\beta X_i+u_i)-(\alpha+\beta\bar{X}+\bar{u})=\beta(X_i-\bar{X})+(u_i-\bar{u})$$

$$y_i=\beta x_i+(u_i-\bar{u})$$

از طرف دیگر، چون $\bar{Y}=\bar{\hat{Y}}$ است، خطای تخمین را به‌صورت زیر می‌نویسیم:

$$e_i=Y_i-\hat{Y}_i=(Y_i-\bar{Y})-(\hat{Y}_i-\bar{Y})=y_i-\hat{y}_i$$

$$=\beta x_i+(u_i-\bar{u})-\hat{\beta}x_i=-(\hat{\beta}-\beta)x_i+(u_i-\bar{u})$$

مجموع مجذور خطاها برابر است با:

$$\sum e_i^۲ = \sum\left(-(\hat{\beta}-\beta)x_i + (u_i - \bar{u})\right)^۲$$
$$=(\hat{\beta}-\beta)^۲ \sum x_i^۲ + \sum (u_i - \bar{u})^۲ - ۲(\hat{\beta}-\beta)\sum x_i(u_i - \bar{u})$$

امید ریاضی $\sum e_i^۲$ برابر است با (در جمله سوم از رابطه $\hat{\beta}-\beta=\sum x_i u_i$ و $\sum x_i(u_i-\bar{u})=\sum x_i u_i$ استفاده شده است).

$$E(\sum e_i^۲)=\sum x_i^۲ E(\hat{\beta}-\beta)^۲ + E\left[\sum (u_i-\bar{u})^۲\right] + ۲E\left[(\sum w_i u_i)(\sum x_i u_i)\right] \qquad (۲-۱۸)$$

حال امید ریاضی هر یک از عبارت‌های فوق را حساب می‌کنیم:

$$E(\hat{\beta}-\beta)=\text{var}(\hat{\beta})=\frac{\sigma^۲}{\sum x_i^۲}$$

$$E\left[\sum (u_i-\bar{u})^۲\right]=\sigma^۲ E\left[\sum \left(\frac{u_i-\bar{u}}{\sigma}\right)^۲\right]=\sigma^۲ E(\chi^۲_{n-۱})=\sigma^۲(n-۱)$$

$$E\left[(\sum w_i u_i)(\sum x_i u_i)\right]=E\left(\sum w_i x_i u_i^۲+...\right)=E\left(\sum w_i x_i u_i^۲\right)$$
$$=E\left(\sum \frac{x_i}{\sum x_i^۲} x_i u_i^۲\right)=E\left(\frac{\sum x_i^۲ u_i^۲}{\sum x_i^۲}\right)=\frac{\sum x_i^۲ E(u_i^۲)}{\sum x_i^۲}=\frac{\sum x_i^۲ \sigma^۲}{\sum x_i^۲}=\sigma^۲$$

در محاسبه امید ریاضیِ جملهٔ آخر، امید ریاضی حاصل‌ضرب‌های متقاطع برابر صفر است، زیرا شامل حاصل‌ضرب‌هایی به شکل $u_i u_j$ است که امید ریاضی آنها صفر است (فرض عدم خودهمبستگی). بنابراین، با جایگذاری در (۱۸-۲) خواهیم داشت:

$$E(\sum e_i^۲)=\sum x_i^۲ \frac{\sigma^۲}{\sum x_i^۲}+(n-۱)\sigma^۲-۲\sigma^۲=\sigma^۲+(n-۱)\sigma^۲-\sigma^۲=(n-۲)\sigma^۲ \qquad (۲-۱۹)$$

با تقسیم طرفین بر $n-۲$، خواهیم داشت:

$$\sigma^۲=\frac{E(\sum e_i^۲)}{n-۲}$$

بنابراین تخمین $\sigma^۲$ برابر است با:

$$\hat{\sigma}^۲=\frac{\sum e_i^۲}{n-۲} \qquad (۲-۲۰)$$

بدین ترتیب $\hat{\sigma}^۲$ تخمین زنندهٔ بدون تورش $\sigma^۲$ است، زیرا طبق (۱۹-۲) خواهیم داشت:

$$E(\hat{\sigma}^۲)=E\left(\frac{\sum e_i^۲}{n-۲}\right)=\frac{E(\sum e_i^۲)}{n-۲}=\frac{(n-۲)\sigma^۲}{n-۲}=\sigma^۲ \qquad (۲-۲۱)$$

e_i از رابطه $e_i = Y_i - (\hat{\alpha}+\hat{\beta}X_i)$ محاسبه می‌شود که نیاز به $\hat{\alpha}$ و $\hat{\beta}$ داریم. اما چون برای برآورد $\hat{\alpha}$ و $\hat{\beta}$ از روابط $\sum e_i = ۰$ و $\sum e_i X_i = ۰$ استفاده شده است، لذا به‌دلیل وجود این دو محدودیت، درجه آزادی $\sum e_i$ برابر $n-۲$ خواهد شد. به‌هرحال، $\hat{\sigma}$ انحراف معیار تخمین است و نشان می‌دهد که به‌طور متوسط Yهای واقعی چه‌قدر از Yهای تخمینی $(\hat{Y})$ انحراف دارد.

مثال ٤-۲: در مثال ۲-۲ خطای معادله رگرسیون را محاسبه کنید.

در این مثال مقادیر زیر را داریم:

$$TSS=\sum y_i^۲=۴۰۹/۸$$

$$ESS=\hat{\beta}^۲\sum x_i^۲=(۱/۳۸)^۲(۱۵۸/۸)=۳۰۲/۵۷$$

$$RSS=TSS-ESS=۴۰۹/۸-۳۰۲/۵۸=۱۰۷/۲۳$$

$$\hat{\sigma}^۲=\frac{RSS}{n-۲}=\frac{۱۰۷/۲۳}{۱۸}=۵/۹۵۷$$

$$\hat{\sigma}=\sqrt{۵/۹۵۷}=۲/۴۴۱$$

$\hat{\sigma}=۲/۴۴$ بیانگر آن است که انحراف Y واقعی از تخمینی، به‌طور متوسط ۲/۴۴ است.

۱۱-۲ آزمون معنادار بودن ضرایب رگرسیون

این آزمون بیانگر این است که آیا ضرایب رگرسیون واقعاً از صفر متفاوت هستند یا خیر. با استفاده از مقادیر $\hat{\alpha}$ و $\hat{\beta}$، می‌توان این آزمون‌ها را به‌سادگی انجام داد. این آزمون برای β به‌صورت آزمون فرضیه $H_۰ : \beta = ۰$ در مقابل فرضیه $H_۱ : \beta \neq ۰$ می‌باشد که برای آزمون آن از توزیع t استفاده می‌شود (زیرا $\hat{\beta}$ توزیع نرمال با میانگین β و واریانس $\sigma^۲/\sum x_i^۲$ دارد که در صورت مجهول بودن $\sigma^۲$ تبدیل به توزیع t می‌شود):

$$t_{n-۲}=\left.\frac{\hat{\beta}-\beta}{\sqrt{\text{var}(\hat{\beta})}}\right|_{H_۰}=\frac{\hat{\beta}}{\frac{\hat{\sigma}}{\sqrt{\sum x_i^۲}}}=\sqrt{\sum x_i^۲}\,\frac{\hat{\beta}}{\hat{\sigma}} \qquad (۲-۲۲)$$

اگر قدرمطلق t از tجدول بیشتر باشد، فرضیه $H_۰ : \beta = ۰$ رد می‌شود و بیانگر معنی‌دار بودن ضریب رگرسیون است. به‌طور کلی هر یک از ضرایب معادله رگرسیون در صورتی در سطح ۹۵

درصد معنی‌دار هستند که برای هر یک از آنها (به شرطی که حجم نمونه به اندازه کافی بزرگ (حدود ۳۰) باشد)، تقریباً $|t| \geq ۲$ باشد.

برای آزمون معنی‌دار بودن α نیز می‌توان فرضیه $H_۰ : \alpha = ۰$ را در مقابل $H_۱ : \alpha \neq ۰$ آزمون نمود که در این حالت آماره t عبارت است از:

$$t_{n-۲} = \frac{\hat{\alpha} - \alpha}{\sqrt{\text{var}(\hat{\alpha})}} \Bigg|_{H_۰} = \frac{\hat{\alpha}}{\hat{\sigma}\sqrt{\frac{۱}{n} + \frac{\bar{X}^۲}{\sum x_i^۲}}} \qquad (۲-۲۳)$$

مثال۵-۲: در مثال ۲-۲ آزمون معنی‌دار بودن ضرایب رگرسیون را انجام دهید.

ابتدا $\hat{\sigma}^۲$ را حساب می‌کنیم که در مثال ۴-۲ برابر با ۲/۴۴۱ به‌دست آمد. مقدار t برای آزمون فرضیه $H_۰ : \beta = ۰$ برابر است با:

$$t = \sqrt{۱۵۸/۸}\,\frac{۱/۳۸}{۲/۴۴۱} = ۷/۱۲۷$$

از آنجا که مقدار $t = ۷/۱۲۷$ از t جدول (در اینجا حجم نمونه n=۲۰ است و لذا t جدول برابر با ۲/۰۹ می‌باشد) بزرگتر است و در ناحیه بحرانی قرار گرفته لذا فرضیه $\beta = ۰$ رد می‌شود و درنتیجه β تفاوت معناداری با صفر دارد.

مقدار t برای آزمون $\alpha = ۰$ عبارت است از:

$$t = \frac{۳/۲۶۶}{\sqrt{۲/۴۴۱\left(\frac{۱}{۲۰} + \frac{(۶/۴)^۲}{۱۵۸/۸}\right)}} = ۲/۴۱۱$$

چون مقدار $t = ۲/۴۱$ از t جدول یعنی ۲/۰۹ بزرگتر است و در ناحیه بحرانی قرار گرفته است، لذا فرضیه $\alpha = ۰$ رد می‌شود و در نتیجه، α تفاوت معناداری با صفر دارد.

این آزمون‌ها را می‌توان برای فرضیه‌هایی مانند $\beta = ۱$ یا $\beta = ۵$ و غیره نیز انجام داد. مراحل انجام این آزمون‌ها تفاوتی با آزمون فرضیه $\beta = ۰$ ندارد.

مثال ٦-۲: در مثال ۲-۲ آیا β می‌تواند برابر با ۱ باشد؟

در اینجا فرضیه $\beta = ۱$ را در مقابل $\beta \neq ۱$ آزمون می‌کنیم. مقدار t برابر است با:

$$t = \sqrt{۱۵۸/۸}\,\frac{۱/۳۸ - ۱}{۲/۴۴۱} = ۱/۹۶$$

چون مقدار t در ناحیه بحرانی قرار نگرفته است لذا فرضیه $\beta = ۱$ رد نمی‌شود. بنابراین β تفاوت معناداری با ۱ ندارد.

برای ضرایب α و β می‌توان فاصله اطمینان نیز تشکیل داد که با استفاده از آماره t به‌دست می‌آید. برای آزمون فرضیه در مورد β، ناحیه قبول فرضیه H_0 به‌صورت زیر است:

$$|t| \le t_{\frac{\alpha}{۲},n-۲} \Rightarrow \left|\frac{\hat{\beta}-\beta}{\hat{\sigma}/\sqrt{\sum x_i^۲}}\right| \le t_{\frac{\alpha}{۲},n-۲}$$

با ساده کردن این نامساوی، می‌توان تخمین فاصله‌ای را برای β و مشابه آن را برای α به‌دست آورد:

$$\left(\hat{\beta}-t_{\frac{\alpha}{۲},n-۲}\frac{\hat{\sigma}}{\sqrt{\sum x_i^۲}} \le \beta \le \hat{\beta}+t_{\frac{\alpha}{۲},n-۲}\frac{\hat{\sigma}}{\sqrt{\sum x_i^۲}}\right)$$

(۲-۲۴)

$$\left(\hat{\alpha}-t_{\frac{\alpha}{۲},n-۲}\,\hat{\sigma}\sqrt{\frac{۱}{n}+\frac{\overline{X}^۲}{\sum x_i^۲}} \le \alpha \le \hat{\alpha}+t_{\frac{\alpha}{۲},n-۲}\,\hat{\sigma}\sqrt{\frac{۱}{n}+\frac{\overline{X}^۲}{\sum x_i^۲}}\right)$$

مثال۷-۲: تخمین فاصله‌ای ۹۵ درصد را برای β در مثال ۲-۲ به‌دست آورید.

محاسبات لازم برای تعیین فاصله اطمینان، در مثال ۲-۲ تا ۵-۲ انجام شده است که عبارتند از:

$$\hat{\beta}=۱/۳۸ \qquad \sum x_i^۲ = ۱۵۸/۸$$

$$\hat{\sigma}^۲ = ۵/۹۵۷ \qquad n = ۲۰$$

با توجه به $t_{/۰۲۵,۱۸} = ۲/۱$ و انجام محاسبات لازم، تخمین فاصله‌ای برای β عبارت است از:

(۰/۹۷۸ , ۱/۷۸۲)

بنابراین با احتمال ۹۵ درصد β در این فاصله قرار دارد.

۲-۱۲ تحلیل واریانس (آزمون معنادار بودن رگرسیون)

تحلیل واریانس روش دیگری برای بررسی وضعیت مدل رگرسیون است. در تحلیل واریانس، عوامل مؤثر بر Y به دو دسته تقسیم می‌شوند. یکی ناشی از X (جزء غیرتصادفی یا قابل کنترل) که برابر با $\alpha+\beta X$ است و دیگری ناشی از اثر سایر عوامل یعنی u (جزء تصادفی یا غیر قابل کنترل) می‌باشد. حال سؤال این است که آیا اثرات ناشی از X یا عاملی که قابل کنترل است در

مقایسه با سایر عوامل یعنی u در خور توجه است یا خیر. این سؤال معادل با این است که اگر $\beta = ۰$ باشد، بدان معنا است که X بی‌تأثیر است و تغییرات Y عملاً ناشی از تغییرات تصادفی است. لذا بحث فوق معادل است با آزمون این فرضیه که آیا β تفاوت معناداری با صفر دارد یا خیر.[۱]

همان‌طور که اشاره شد، کل تغییرات Y برابر با «تغییرات توضیح داده شده» به‌علاوهٔ «تغییرات توضیح داده نشده» می‌باشد که اولی ناشی از X و دومی ناشی از u است:

$$TSS=ESS+RSS$$

$$\sum(Y_i-\overline{Y})^۲=\sum(\hat{Y}_i-\overline{Y})^۲+\sum e_i^۲ \qquad (۲-۲۵)$$

اما هر یک از جملات رابطه (۲-۲۵) را می‌توان برحسب توزیع $\chi^۲$ نوشت. زیرا شامل متغیرهایی هستند که دارای توزیع نرمال بوده (Y_i و $\hat{Y}_i$ هر کدام توزیع نرمال دارند، زیرا تابع خطی از u_i هستند) و به‌صورت مجذور نیز بیان شده‌اند.

$$TSS=\sum y_i^۲ \quad , \quad ESS=\sum \hat{y}_i^۲ \quad , \quad RSS=\sum e_i^۲=\sum(Y_i-\hat{Y}_i)^۲$$

TSS شامل n متغیر Y_i می‌باشد. اما به‌دلیل محدودیت $\sum(Y_i-\overline{Y})=۰$ درجه آزادی آن $n-۱$ می‌باشد. ESS تابعی از $\hat{Y}_i$ است که آن نیز فقط تابعی از دو متغیر تصادفی $\hat{\alpha}$ و $\hat{\beta}$ است و به‌دلیل اینکه $\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{X}$ است لذا درجه آزادی آن برابر با ۱ خواهد شد. RSS نیز تابعی از n متغیر تصادفی e_i است. اما چون دو محدودیت به‌صورت $\sum e_i=۰$ و $\sum e_iX_i=۰$ دارد، لذا درجه آزادی آن $n-۲$ است.

حال می‌توان ESS (یعنی اثر X) را با RSS (اثر سایر عوامل یا u) مقایسه نمود. بدین منظور ابتدا میانگین آنها را محاسبه کرده و سپس از نسبت آنها، تابع F را تعریف می‌کنیم[۲]:

۱- در فصل سوم خواهیم دید که در رگرسیون چند متغیره، آزمون معنادار بودن رگرسیون با آزمون معنی‌دار بودن ضریب β، متفاوت است.

[2] توجه شود که طبق رابطه (۱-۳۰) در فصل اول، روابط زیر برقرار است:

$$\frac{TSS}{\sigma^۲}=\frac{\sum y_i^۲}{\sigma^۲}=\chi^۲_{n-۱} \quad , \quad \frac{ESS}{\sigma^۲}=\frac{\sum \hat{y}_i^۲}{\sigma^۲}=\chi^۲_۱ \quad , \quad \frac{RSS}{\sigma^۲}=\frac{\sum e_i^۲}{\sigma^۲}=\chi^۲_{n-۲}$$

$$S_E^2 = \frac{ESS}{1} \quad , \quad S_R^2 = \hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum e_i^2}{n-2}$$

$$F_{1,n-2} = \frac{\frac{\chi_1^2}{1}}{\frac{\chi_{n-2}^2}{n-2}} = \frac{\frac{\sigma^2\chi_1^2}{1}}{\frac{\sigma^2\chi_{n-2}^2}{n-2}} = \frac{S_E^2}{S_R^2} = \frac{ESS}{\frac{RSS}{n-2}} = (n-2)\frac{ESS}{RSS} \qquad (2-26)$$

بدیهی است که اگر مقدار F بزرگ باشد، نشان می‌دهد که معادله رگرسیون توانسته است به‌خوبی تغییرات Y را توضیح دهد و از نظر آماری معنادار است، زیرا بزرگ بودن F بدین معنی است که تأثیر X در مقایسه با سایر عوامل (u) درخور توجه است.

از طرف دیگر می‌توان F را بر حسب R^2 نیز بیان نمود:

$$\begin{aligned} F &= (n-2)\frac{ESS}{RSS} = (n-2)\frac{ESS}{TSS-ESS} \\ &= (n-2)\frac{ESS/TSS}{1-ESS/TSS} = (n-2)\frac{R^2}{1-R^2} \end{aligned} \qquad (2-27)$$

بنابراین هر چقدر R^2 بیشتر باشد، مقدار F نیز بیشتر خواهد شد. به هر حال مزیت آزمون F در این است که با استفاده از آن می‌توان معناداربودن رگرسیون‌های چندمتغیره را نیز آزمون نمود.[1]

مثال ۱۰-۲: در مثال ۲-۲ آیا رگرسیون Y روی X معنادار است یا خیر.

چون $ESS = ۱۵/۵۵۸$ و $TSS = ۲۴/۸۳۳$ لذا $RSS = ۲۴/۸۳۳ - ۱۵/۵۵۸ = ۹/۲۷۵$ است. بر این اساس مقدار F برابر است با:

$$F = (n-2)\frac{ESS}{RSS} = (20-2)\frac{302/57}{107/23} = 50/79$$

$$\left(F \geq_{\alpha,1,n-2} = F_{0/05,1,18} = 8/29\right)$$

چون $F = ۵۰/۷۹$ از F جدول (۸/۲۹) بزرگتر است، لذا فرضیه H_0 رد می‌شود. به‌عبارت دیگر X توانسته است به‌طور معناداری تغییرات Y را توضیح دهد.

---

[1] توجه شود که در رگرسیون یک متغیره، چون درجه آزادی صورت F برابر با ۱ است، لذا F برابر بامجذور t در آزمون فرضیه $H_0 : \beta = 0$ است.

۲-۱۳ جمع‌بندی و تحلیل نتایج رگرسیون

در تحلیل رگرسیون، متناسب با موضوع مورد نظر، ابتدا معادله را انتخاب می‌کنیم که ممکن است خطی یا غیرخطی باشد. سپس ضرایب آن را با روش حداقل مربعات معمولی (OLS) برآورد می‌کنیم. برای این معادله، معیارهایی از قبیل معنی‌دار بودن ضرایب، ضریب تعیین و مقدار F محاسبه می‌شوند که همراه با هر معادلهٔ رگرسیون، ارائه می‌گردند. در مثال زیر، این معیارها و تفسیر مختصر آنها ارائه شده است.

مثال ۱۱-۲: نتایج حاصل از مثال ۲-۲ به‌طور خلاصه عبارت است از:

$$\hat{Y}_i = ۳/۲۶۶ + ۱/۳۸X_i \qquad R^۲ = ۰/۷۳۸$$
$$(۲/۴) \quad (۷/۱) \qquad F = ۵۰/۷۹$$

$R^۲$ بیانگر این است که معادلهٔ تخمینی توانسته است حدود ۷۳/۸ درصد از تغییرات متغیر وابسته (Y) را توضیح دهد ولی ۲۶/۲ درصد از تغییرات Y ناشی از عوامل تصادفی است که این معادله نتوانسته آنها را به حساب آورد. به‌عبارت دیگر، ۷۳/۸ درصد از تغییرات Y توسط X توضیح داده می‌شود و بقیه تغییرات آن ناشی از سایر عوامل است که در داخل u قرار دارند. مقدار F نیز نشان می‌دهد که این معادله کاملاً معنی‌دار است زیرا از F جدول بیشتر می‌باشد.

ارقام داخل پرانتز بیانگر مقدار t برای هر یک از این ضرایب می‌باشد که وضعیت معنی‌دار بودن آنها را نشان می‌دهد. مقدار t برای β بیشتر از ۲ است، لذا معنی‌دار می‌باشد (به‌عبارت دیگر به‌طور معنی‌داری از صفر تفاوت دارد). همچنین چون مقدار t برای α بیشتر از ۲ است، لذا معنی‌دار بوده و با صفر تفاوت دارد.

برآورد رگرسیون ساده با Eviews

فایل **data1**

برای برآورد معادله رگرسیون، به دو صورت می‌توان عمل نمود. یک روش این است که در سطر فرمان با فرمان LS، معادله مورد نظر را به‌صورت LS Y C X وارد نماییم. LS فرمان اجرای روش حداقل مربعات معمولی است. در اینجا Y متغیر وابسته، C عرض از مبدأ و X متغیر توضیحی را نشان می‌دهد. بعد از وارد نمودن عبارت مذکور با زدن کلید Enter نتایج تخمین در پنجره Equation ارائه می‌شود.

همچنین می‌توان از منوی Quick گزینه Estimation Equation را انتخاب نمود که به‌دنبال آن پنجره زیر باز می‌شود:

Equation Estimation

Specification | Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1345 1380

OK | Cancel

در اینجا معادله مورد نظر را به‌صورت Y C X و یا به صورت Y=C(1)+C(2)*X وارد نموده و OK را انتخاب می‌کنیم. در قسمت پایین این پنجره، روش‌های تخمین نیز مشخص شده است که در گزینه Method قرار دارند. گزینه Sample نیز بیانگر دوره‌ای است که می‌خواهیم معادله را برای آن دوره برآورد نماییم.

به‌عنوان نمونه، نتایج تخمین معادله سرمایه‌گذاری بخش خصوصی به قیمت ثابت (IP) روی نرخ بهره حقیقی (RR) برای دوره ۶۵-۱۳۴۵ به‌صورت زیر می‌باشد.

Equation: UNTITLED Workfile: S1::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: IP
Method: Least Squares
Date: 02/03/11 Time: 08:38
Sample: 1345 1365
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20131.31	2587.601	7.779914	0.0000
RR	-844.9209	324.5470	-2.603385	0.0175

R-squared	0.262926	Mean dependent var	22847.13
Adjusted R-squared	0.224133	S.D. dependent var	12319.65
S.E. of regression	10851.55	Akaike info criterion	21.51240
Sum squared resid	2.24E+09	Schwarz criterion	21.61188
Log likelihood	-223.8802	Hannan-Quinn criter.	21.53399
F-statistic	6.777613	Durbin-Watson stat	1.177749
Prob(F-statistic)	0.017459		

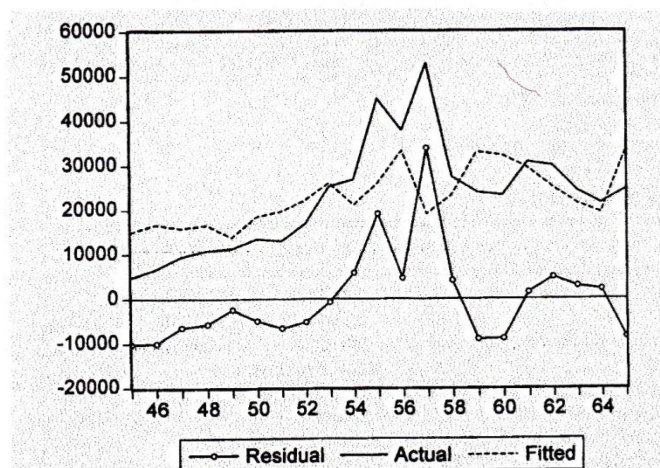
همان‌طور که ملاحظه می‌شود در بالای این جدول، نام متغیر وابسته مشخص شده است. سپس روش تخمین با عبارت Least Squares نشان داده شده است. در سطر بعدی، زمان انجام این تخمین و سپس دوره زمانی در مقابل Sample ارائه شده است. تعداد مشاهدات نیز مشخص شده که برابر با ۲۱ می‌باشد.

در قسمت بعد ابتدا ستون Variable را داریم که در این ستون نام متغیرهای توضیحی ارائه می‌شود. C بیانگر عرض از مبدأ و RR نیز نرخ بهره حقیقی است. در ستون دوم که تحت عنوان Coefficient است مقدار ضرایب را نشان می‌دهد که مقدار عرض از مبدأ ۲۰۱۳۱/۳۱ و ضریب نرخ بهره یا شیب معادله رگرسیون برابر با ۸۴٤/۹- می‌باشد. در ستون بعدی، انحراف معیار تخمین‌زننده‌ها (Std. Error) ارائه شده است. همچنین از تقسیم مقدار هر ضریب بر انحراف معیار آن، مقدار آمارهٔ t به‌دست می‌آید که در ستـــون t-Statistic ارائه شده است. چون قدرمطلق t برای هر یک از ضرایب بیشتر از ۲ است لذا هر دو ضریب معنی‌دار هستند. در ستون آخر یعنی .Prob سطح معنی‌دار بودن ضرایب را نشان می‌دهد که اگر کمتر از ۰/۰۵ باشد بدین معنی است که ضرایب، لااقل در سطح ۰/۰۵ معنی‌دار هستند.

در قسمت پایین جدول، مقدار ضریب تعیین (R^2) با R-Squared و مقدار F نیز با F-Statistic مشخص شده‌اند. در این مثال R^2 =۰/۲٦۳ و F=٦/۷۸ می‌باشد. همچنین در زیر تابع F سطح معنی‌دار بودن F (یا معنی‌دار بودن معادله رگرسیون) ارائه شده است که در اینجا ۰/۰۱۷ می‌باشد. به‌طور کلی اگر این احتمال از ۰/۰۵ کمتر باشد، بدین معنی است که معادله رگرسیون معنادار است.

همچنین عبارت Mean dependent var بیانگر میانگین متغیر وابسته و S.D. dependent var نیز بیانگر انحراف معیار متغیر وابسته است. علاوه بر این، S.E. of regresion نیز انحراف معیار معادله رگرسیون یعنی $\hat{\sigma}$ می‌باشد. مجموع مجذور خطاها یا RSS نیز با Sum squared resid نشان داده شده است. راجع به بقیه ارقامی که در جدول مذکور ارائه شده است در ادامه و به‌ویژه در فصل‌های بعدی بحث خواهد شد.

در پنجره Equation که نتایج معادله رگرسیون را نشان می‌دهد، منوهای مختلفی وجود دارد که هر کدام کارکردهای خاص خود را دارند. به‌عنوان مثال در منوی View با انتخاب Actual, Fitted, Residual می‌توان نمودار و جدول مقادیر واقعی متغیر وابسته (Actual)، مقادیر برآوردی (Fitted) و باقیمانده‌ها (Residual) را به‌دست آورد. برای تابع سرمایه‌گذاری نمودار زیر به‌دست آمده است.



همچنین در منوی View گزینه‌های دیگری جهت انجام آزمون‌های مختلف وجود دارد که در قسمت‌های بعدی به آنها اشاره خواهد شد.

۱۴-۲ پیش‌بینی و فاصله اطمینانِ پیش‌بینی

پیش‌بینی یکی از ابزارهای مهم برای تصمیم‌گیرهای اقتصادی و مالی است و لذا یکی از مهمترین اهداف مدل‌های اقتصاد سنجی می‌باشد. دقت این پیش‌بینی‌ها می‌تواند منجر به تصمیمات بهتر و دقیق‌تر شود و منافع آنها را بیشتر نماید.

پیش‌بینی‌ها را می‌توان بر اساس دوره زمانی، به دو دسته تقسیم نمود: پیش‌بینی درون‌نمونه‌ای و پیش‌بینی برون‌نمونه‌ای. پیش‌بینی درون‌نمونه‌ای مربوط به دوره‌ای است که مدل را برای همان دوره برآورد نموده‌ایم. به‌عنوان مثال اگر مدل برای دوره ۸۰-۱۳۵۰ برآورد شده باشد، در این صورت پیش‌بینی درون‌نمونه‌ای نیز برای دوره ۸۰-۱۳۵۰ انجام می‌شود. معمولاً نتایج پیش‌بینی درون‌نمونه‌ای نسبتاً خوب هستند. این نوع پیش‌بینی معمولاً برای بررسی دقت مدل می‌باشد، زیرا می‌توان مقادیر پیش‌بینی را با مقادیر واقعی مقایسه نمود.

پیش‌بینی برون‌نمونه‌ای، برای دورهٔ بعد از برآورد مدل می‌باشد. مثلاً اگر دورهٔ برآورد شامل ۸۰-۱۳۵۰ است، دورهٔ بعد از آن می‌تواند شامل ۸۵-۱۳۸۱ یا ۹۵-۱۳۸۱ باشد. اگر در پیش‌بینی برون‌نمونه‌ای، دورهٔ پیش‌بینی مربوط به گذشته باشد در این صورت امکان مقایسه مقادیر پیش‌بینی با مقادیر واقعی وجود دارد. ولی اگر مربوط به دوره آینده باشد امکان مقایسه وجود ندارد، زیرا مقادیر واقعی هنوز تحقق نیافته‌اند.

برای تعیین دقت پیش‌بینی، معیارهای مختلفی وجود دارد که در فصل یازدهم ارائه شده است.

با استفاده از معادله رگرسیون می‌توان مقادیر Y را پیش‌بینی نمود. تخمین مقدار پیش‌بینی‌شدهٔ Y را با $\hat{Y}_f$ نشان می‌دهیم. به‌عبارت دیگر اگر مقدار X در سال f برابر با X_f باشد آنگاه پیش‌بینی ما از Y برابر با $\hat{Y}_f$ است:

$$\hat{Y}_f = \hat{\alpha} + \hat{\beta} X_f$$

اما مقدار دقیق Y در سال f برابر خواهد بود با:

$$Y_f = \hat{Y}_f + e_f$$

e_f مقدار خطای پیش‌بینی است. در اینجا نیز برای ساختن تخمین فاصله‌ای برای Y_f، ابتدا امید ریاضی و واریانس خطای پیش‌بینی را محاسبه می‌کنیم:

$$e_f = Y_f - \hat{Y}_f \quad ; Y_f = \alpha + \beta X_f + u_f \quad , \quad \hat{Y}_f = \hat{\alpha} + \hat{\beta} X_f \tag{۲-۲۸}$$

توجه شود که خطای واقعیِ پیش‌بینی برابر با u_f است که تخمین آن برابر با e_f می‌باشد.

$$\begin{aligned} \text{var}(e_f) &= \text{var}(Y_f - \hat{Y}_f) \\ &= \text{var}(Y_f) + \text{var}(\hat{Y}_f) - ۲\text{cov}(Y_f, \hat{Y}_f) = \text{var}(Y_f) + \text{var}(\hat{Y}_f) \end{aligned} \tag{۲-۲۹}$$

Y_f و $\hat{Y}_f$ مستقل از هم هستند، زیرا Y_f تابعی از u_f است اما $\hat{Y}_f$ تابعی از $\hat{\alpha}$ و $\hat{\beta}$ بوده که این دو نیز تابعی از u_i هستند. چون طبق فرض «عدم خودهمبستگی»، u_i و u_f مستقل از هم هستند، لذا باعث می‌شود که Y_f و $\hat{Y}_f$ نیز مستقل از هم باشند و کوواریانس آنها صفر گردد:

$$\text{var}(e_f) = \text{var}(Y_f) + \text{var}(\hat{Y}_f) = \text{var}(u_f) + \text{var}(\hat{\alpha} + \hat{\beta} X_f)$$

زیرا طبق (۲-۲۸)، $\text{var}(Y_f) = \text{var}(u_f) = \sigma^۲$ است. با محاسبه $\text{var}(\hat{\alpha} + \hat{\beta} X_f)$ خواهیم داشت:[1]

$$\text{var}(e_f) = \sigma^۲ + \sigma^۲ \left(\frac{۱}{n} + \frac{(X_f - \bar{X})^۲}{\sum x_i^۲} \right) = \sigma^۲ \left(۱ + \frac{۱}{n} + \frac{(X_f - \bar{X})^۲}{\sum x_i^۲} \right) \tag{۲-۳۰}$$

حال تابع t را تشکیل می‌دهیم:

$$t_{n-۲} = \frac{e_f - E(e_f)}{\sqrt{\text{var}(e_f)}} = \frac{e_f - ۰}{\sqrt{\text{var}(e_f)}} = \frac{Y_f - \hat{Y}_f}{\sqrt{\text{var}(e_f)}}$$

با استفاده از شرط $|t| \le t_{\frac{\alpha}{۲}, n-۲}$، تخمین فاصله‌ای برای Y_f عبارت است از:

$$\begin{aligned} &\left(\hat{Y}_f - t_{\frac{\alpha}{۲}, n-۲} \sqrt{\text{var}(e_f)} \le Y_f \le \hat{Y}_f + t_{\frac{\alpha}{۲}, n-۲} \sqrt{\text{var}(e_f)} \right) \\ &\text{var}(e_f) = \hat{\sigma}^۲ \left(۱ + \frac{۱}{n} + \frac{(X_f - \bar{X})^۲}{\sum x_i^۲} \right) \end{aligned} \tag{۲-۳۱}$$

---

[1] برای محاسبهٔ واریانس $\hat{\alpha} + \hat{\beta} X_f$، ابتدا به‌جای $\hat{\alpha}$ قرار داده و سپس آن را به‌صورت زیر حساب می‌کنیم:

$$\text{var}(\hat{\alpha} + \hat{\beta} X_f) = \text{var}(\bar{Y} - \hat{\beta}\bar{X} + \hat{\beta} X_f) = \text{var}[\bar{Y} + \hat{\beta}(X_f - \bar{X})]$$

$$= \text{var}(\bar{Y}) + (X_f - \bar{X})^۲ \text{var}(\hat{\beta}) = \frac{\sigma^۲}{n} + \frac{(X_f - \bar{X})^۲ \ \sigma^۲}{\sum (X_i - \bar{X})^۲}$$

مثال ۱۳-۲: در مثال ۱۲-۲ اگر مقدار $X_f = ۱۰$ باشد، مقدار پیش‌بینی شدهٔ Y_f را محاسبه کرده و تخمین فاصله را برای آن به‌دست آورید.

مقدار پیش‌بینی شدهٔ Y_f برابر با $\hat{Y}_f$ است:

$$\hat{Y}_f = ۳/۲۶۶ + ۱/۳۸(۱۰) = ۱۷/۰۶۶$$

و با توجه به محاسبات قبلی، تخمین فاصله‌ای به‌ازای $X_f = ۸$ برابر است با:

$$۱۷/۰۶۶ \pm (۲/۱)(۲/۴۴۱)\sqrt{۱ + \frac{۱}{۲۰} + \frac{(۱۰ - ۶/۴)^۲}{۱۵۸/۸}}$$

$$\Rightarrow ۱۷/۰۶۶ \pm ۵/۴۵۳ \Rightarrow (۱۱/۶۱۳ \;,\; ۲۲/۵۱۹)$$

بدین ترتیب با احتمال ۹۵ درصد، مقدار Y_f در فاصله مذکور قرار خواهد داشت.

پیش‌بینی با Eviews

برای انجام پیش‌بینی با Eviews فرض کنید دوره زمانی شامل ۸۵-۱۳۵۰باشد. ابتدا معادله مورد نظر را برای یک دوره دلخواه مانند ۸۰-۱۳۵۰، برآورد می‌کنیم. سپس در پنجرهٔ نتایج، با انتخاب Proc → Forecast پنجره زیر را باز می‌کنیم:

Forecast

Forecast of
Equation: UNTITLED   Series: Y

Series names
Forecast name: yf
S.E. (optional):
GARCH(optional):

Method
Static forecast
(no dynamics in equation)
☐ Structural (ignore ARMA)
☑ Coef uncertainty in S.E. calc

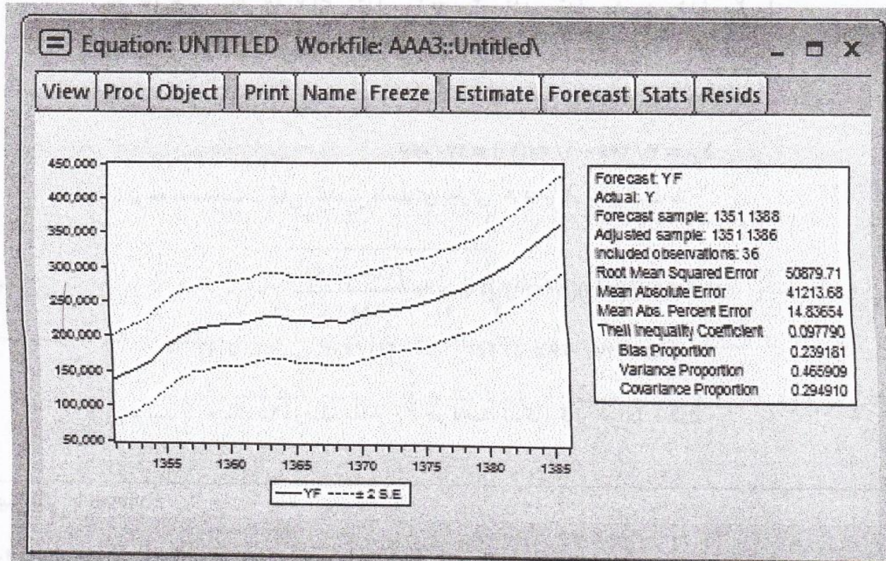
Forecast sample
1350 1385

Output
☑ Forecast graph
☑ Forecast evaluation

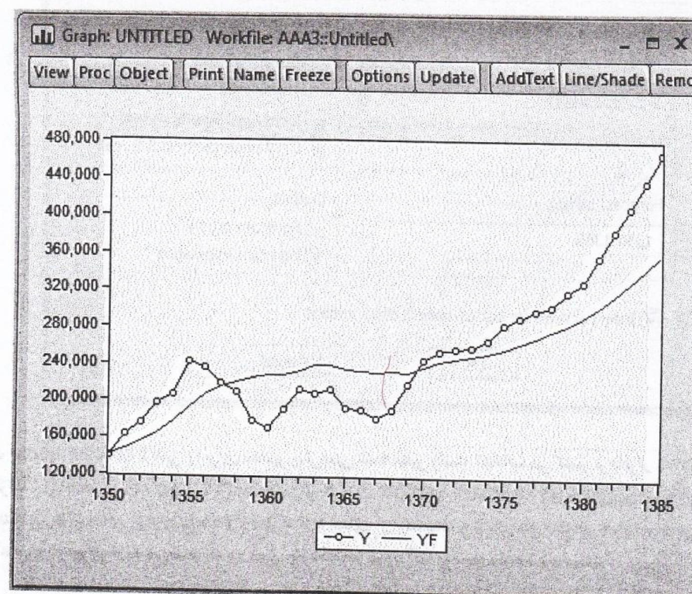
☑ Insert actuals for out-of-sample observations

OK   Cancel

در قسمت Forecast name نامی را برای مقادیر پیش‌بینی شدهٔ متغیر وابسته انتخاب می‌کنیم. نرم‌افزار Eviews به آخر نام متغیر وابسته، حرف f را اضافه می‌کند (مانند Yf). همچنین در قسمت Forecast sample دورهٔ پیش‌بینی را وارد می‌کنیم (مثلاً ۸۵-۱۳۵۰). با انتخاب OK مقادیر پیش‌بینی شدهٔ y همراه با مرزهای آن (به‌اضافه و منهای دو انحراف معیار) ارائه می‌شود. در سمت راست آن اطلاعات دیگری نیز ارائه می‌شود که در فصل یازدهم در مورد آنها توضیح داده شده است.



همچنین می‌توان با فرمان plot Y Yf مقادیر واقعی و پیش‌بینی را در یک نمودار ترسیم کرده و مقایسه نمود. البته می‌توان برای دورهٔ بعد از ۱۳۸۵، مثلاً تا ۱۳۹۰ نیز پیش‌بینی را انجام داد. در این حالت باید Range را تا ۱۳۹۰ گسترش دهیم. از طرف دیگر چون مقادیر X را برای سالهای ۱۳۸۶ تا ۱۳۹۰ نداریم، لذا بایستی مقادیر آن را به‌طور دلخواه یا در قالب سناریوهای مختلف تعیین کرده و سپس Y را تا ۱۳۹۰ پیش‌بینی نمود.



۲-۱۵ رگرسیون‌های غیرخطی

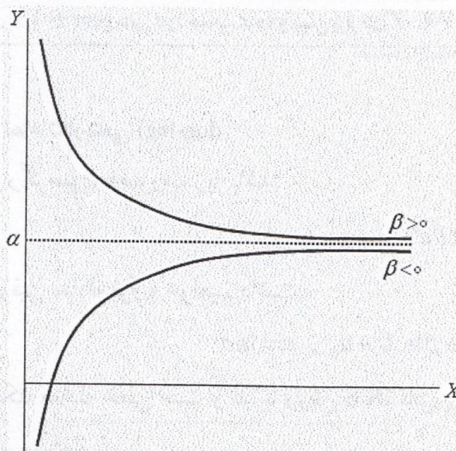
تاکنون معادله رگرسیون خطی را به‌صورت $Y_i=\alpha+\beta X_i+u_i$ در نظر گرفتیم. اما ممکن است رابطه X و Y غیرخطی باشد. در اینجا انواع خاصی از رگرسیون‌های غیرخطی را بررسی می‌کنیم که امکان تبدیل آنها به یک رابطه خطی وجود دارد تا امکان استفاده از روش OLS فراهم شود.

۲-۱۵-۱ روابط معکوس

روابط معکوس در مواردی وجود دارد که با افزایش X مقدار Y به سمت یک مقدار معین میل می‌کند. یکی از توابعی که می‌تواند چنین وضعیتی را نشان دهد عبارت است از:

$$Y_i=\alpha+\frac{\beta}{X_i} \tag{۲-۳۷}$$

نزولی یا صعودی بودن این رابطه، بستگی به علامت β دارد که نمودار زیر آن را نشان می‌دهد.



نمودار ۲-۷: روابط معکوس

برای برآورد این معادله می‌توان آن را به‌صورت خطی نوشت:

$$Y_i=\alpha+\beta X_i^* \ , \ X_i^*=\frac{۱}{X_i}$$

اثر تغییر X بر Y عبارت است از:

$$\frac{dY}{dX}=-\frac{\beta}{X^۲}$$

اگر اثر تغییر X بر Y را بر حسب درصد بیان کنیم بیانگر کشش Y نسبت‌به X است که نشان می‌دهد یک درصد تغییر در X چند درصد Y را تغییر می‌دهد:

$$\eta=\frac{dy}{dx}\frac{X}{Y}=(-\frac{\beta}{X^{2}})\frac{X}{Y}=-\frac{\beta}{XY}$$

برای محاسبه η معمولاً از میانگین‌های X و Y استفاده می‌شود.

$$\eta=-\frac{\beta}{\overline{X}\overline{Y}}$$

برآورد روابط معکوس با Eviews

بدین منظور می‌توان معادله $Y_i=\alpha+\frac{\beta}{X_i}$ را با فرمان زیر برآورد کرد:

```
LS Y C 1/X
```

و یا می‌توان ابتدا با فرمان genr z=1/x متغیر z را حساب کرده و سپس با فرمان LS Y C Z معادله را برآورد نمود.

۲-۱۵-۲ معادلات تمام لگاریتمی (log-log)

فرض کنید معادله رگرسیون به‌صورت زیر باشد:

$$Y_i=aX_i^{\beta}e^{u_i} \quad (۲-۳۸)$$

با گرفتن لگاریتم از طرفین معادله فوق، خواهیم داشت:

$$\ln Y_i=\alpha+\beta\ln X_i+u_i \;\;,\;\; \alpha=\ln a \quad (۲-۳۹)$$

معادله فوق شبیه یک معادله خطی است و لذا با روش OLS قابل برآورد است. اثر متغیر X بر Y به‌صورت زیر حساب می‌شود.

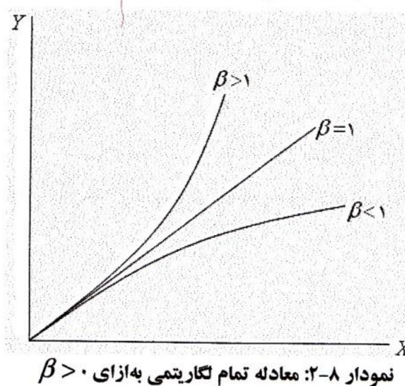
$$\eta=\frac{d\ln Y}{d\ln X}=\frac{dY/Y}{dX/X}=\frac{dY}{dX}\frac{X}{Y}=\beta$$

بنابراین β برابر با کشش Y نسبت‌به X است. برای محاسبه $\frac{dY}{dX}$، رابطه فوق را به‌صورت زیر می‌نویسیم:

$$\frac{dY}{dX}=\beta\frac{Y}{X}$$

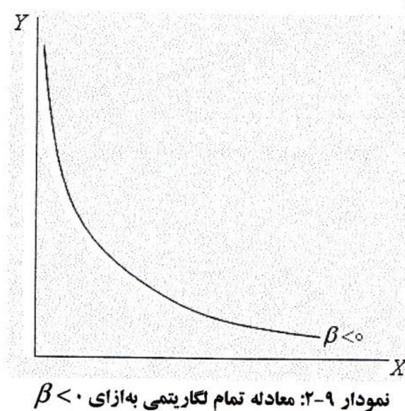
در اینجا نیز برای محاسبه اثر X بر Y معمولاً از میانگین‌های X و Y استفاده می‌شود.

بدیهی است که معادله $Y=ax^{\beta}$ کاملاً انعطاف‌پذیر است. به‌عنوان مثال اگر α و β مثبت باشند، نمودار زیر، رابطه X و Y را توصیف می‌کند.



نمودار ۸-۲: معادله تمام لگاریتمی به‌ازای $\beta > 0$

اگر $\beta < 0$ باشد، خواهیم داشت:



نمودار ۹-۲: معادله تمام لگاریتمی به‌ازای $\beta < 0$

برآورد معادله لگاریتمی با Eviews

بدین منظور فرمان زیر را اجرا می‌کنیم:

LS ln(Y) C ln(X)

که C مقدار α یا $\ln a$ و ضریب ln(X) نیز برابر با β است. بدیهی است که برای تعیین α از رابطه $\ln a=\alpha$ یا $a=e^{\alpha}$ استفاده می‌کنیم.

همچنین می‌توان ابتدا با فرمان genr متغیرهای LY=ln(Y) و LX=ln(X) را محاسبه کرده و سپس معادله رگرسیون لگاریتمی را با فرمان زیر برآورد نمود:

LS LY C LX

۳-۱۵-۲ توابع نمایی

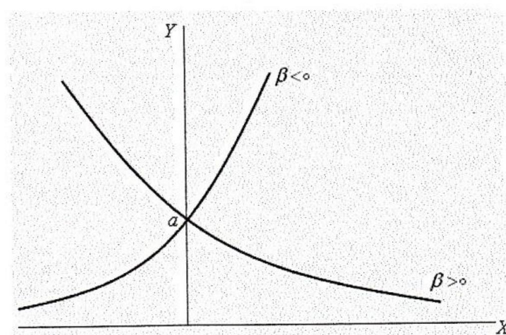
اگر رابطه Y و X توسط یک تابع نمایی توصیف شود، شکل کلی آن به‌صورت زیر می‌باشد.

$$Y_i = ae^{\beta X_i + u_i} = ae^{\beta X_i} e^{u_i} \qquad (۲-۴۰)$$

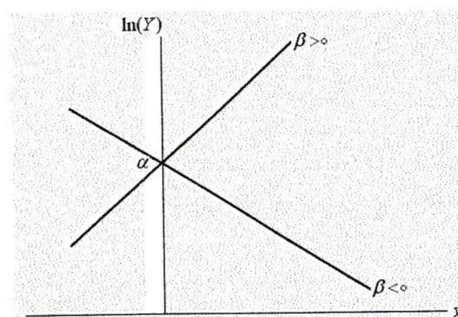
با گرفتن لگاریتم خواهیم داشت:

$$\ln Y_i = \alpha + \beta X_i + u_i \ , \ \alpha = \ln a$$

معادله فوق را log-lin می‌گویند که به معنی لگاریتمی- خطی است، زیرا نسبت‌به Y لگاریتمی و نسبت‌به X خطی است.



نمودار ۱۰-۲: معادله لگاریتمی- خطی



نمودار ۱۱-۲: معادله لگاریتمی- خطی

اثر تغییرات X بر Y برابر است با:

$$\frac{d\ln Y}{dX} = \beta \Rightarrow \frac{dY/Y}{dX} = \beta \Rightarrow \frac{dY}{dX} = \beta Y$$

و کشش Y نسبت‌به X برابر است با:

$$\eta = \frac{dY}{dX}\frac{X}{Y} = \beta Y \frac{X}{Y} = \beta X$$

برآورد تابع نمایی با Eviews

معادله log-lin را با فرمان زیر برآورد می‌کنیم:

```
LS ln(Y) C X
```

C برآورد α و ضریب X برآورد β را نشان می‌دهند. برای محاسبه a از رابطه $\ln a=\alpha$ یا $a=e^{\alpha}$ استفاده می‌کنیم.

۲-۱۵-٤ رگرسیون انحراف از میانگین

معادله رگرسیون خطی به‌صورت زیر است:

$$Y_i=\alpha+\beta X_i+u_i \tag{۲-۴۱}$$

معادله فوق را جمع زده و بر n تقسیم می‌کنیم:

$$\frac{\sum Y_i}{n}=\frac{\sum(\alpha+\beta X_i+u_i)}{n}$$

$$\overline{Y}=\frac{n\alpha+\beta\sum X_i+\sum u_i}{n}\Rightarrow\overline{Y}=\alpha+\beta\overline{X}+\overline{u} \tag{۲-۴۲}$$

معادله (۲-۴۱) را منهای معادله (۲-۴۲) می‌کنیم:

$$Y_i-\overline{Y}=\beta(X_i-\overline{X})+(u_i-\overline{u})$$

انحراف از میانگین را با $Y_i-\overline{Y}=y_i$، $X_i-\overline{X}=x_i$ و $u_i-\overline{u}=v_i$ نشان می‌دهیم:

$$y_i=\beta x_i+v_i \tag{۲-۴۳}$$

بنابراین، اگر داده‌ها را بر حسب انحراف از میانگین بنویسیم، عرض از مبدأ حذف می‌شود، ولی شیب معادله تغییری نمی‌کند. زیرا در رگرسیون $Y_i=\alpha+\beta X_i+u_i$ که تخمین آن به‌صورت $\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i$ است، خطای تخمین برابر است با:

$$\begin{aligned}e_i&=Y_i-\hat{Y}_i=\hat{Y}_i-\hat{\alpha}-\hat{\beta}X_i\\&=Y_i-(\overline{Y}-\hat{\beta}\overline{X})-\hat{\beta}X_i\ ;\ \hat{\alpha}=\hat{Y}-\hat{\beta}\overline{X}\\&=(Y_i-\overline{Y})-\beta(X_i-\overline{X})=y_i-\hat{\beta}x_i\end{aligned} \tag{۲-۴۴}$$

بنابراین با حداقل نمودن مجموع مجذور خطاها، $\hat{\beta}$ به‌دست می‌آید:

$$\sum e_i^۲=\sum(y_i-\hat{\beta}x_i)^۲$$

$$\frac{\partial\sum e_i^۲}{\partial\hat{\beta}}=۲\sum(y_i-\hat{\beta}x_i)(-x_i)=۰\Rightarrow\hat{\beta}=\frac{\sum x_iy_i}{\sum x_۲^۲} \tag{۲-۴۵}$$

بنابراین، تخمین شیب معادله رگرسیون، مشابه قبل است.

برآورد رگرسیون انحراف از میانگین با Eviews

رگرسیون انحراف از میانگین را فرمان زیر برآورد می کنیم:

LS Y-@mean(Y) X-@mean(X)

چون عرض از مبدأ این معادله برابر صفر است، لذا حرف C را ننوشته‌ایم. ضریب X-@mean(X) برابر با تخمین β است.

۵-۱۵-۲ رگرسیون با متغیرهای استاندارد شده

مقادیر استاندارد شدهٔ X و Y عبارتند از:

$$X_i^* = \frac{X_i - \overline{X}}{S_x} = \frac{x_i}{S_x} \ , \ Y_i^* = \frac{Y_i - \overline{Y}}{S_y} = \frac{y_i}{S_y} \qquad (۲-۴۶)$$

توجه شود که $\sum X_i^* = \sum Y_i^* = ۰$ است، زیرا

$$\sum X_i^* = \sum \frac{x_i}{S_x} = \frac{\sum x_i}{S_x} = ۰$$

بنابراین $\overline{X}^* = \overline{Y}^* = ۰$ خواهد بود.

معادله رگرسیون برای داده‌های استانداردشده، عبارت است از:

$$Y_i^* = \alpha^* + \beta^* X_i^* \qquad (۲-۴۷)$$

برای برآورد ضرایب، ابتدا $\sum e_i^{*۲}$ را حساب کرده و سپس نسبت‌به $\hat{\alpha}^*$ و $\hat{\beta}^*$ مشتق می‌گیریم:

$$\sum e_i^{*۲} = \sum (Y_i^* - \hat{\alpha}^* - \hat{\beta}^*)^۲$$

$$\left.\begin{array}{l} \dfrac{\partial \sum e_i^{*۲}}{\partial \hat{\alpha}^*} = ۰ \\ \dfrac{\partial \sum e_i^{*۲}}{\partial \hat{\beta}^*} = ۰ \end{array}\right. \Rightarrow \quad \hat{\alpha}^* = \overline{Y}^* - \hat{\beta}^* \overline{X}^* = ۰ \quad , \quad \hat{\beta}^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*۲}}$$

اگر به‌جای x^* و y^* قرار دهیم، خواهیم داشت:

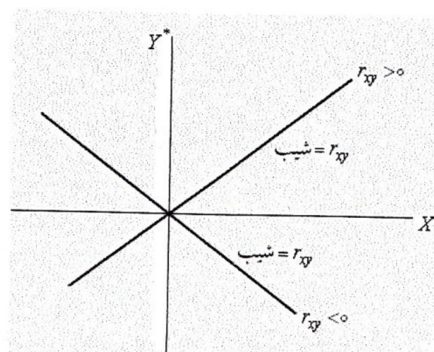
$$\hat{\beta}^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*۲}} = \frac{\sum \frac{x_i}{S_x}\frac{y_i}{S_y}}{\sum \frac{x_i^۲}{S_x^۲}} = \frac{S_x}{S_y}\frac{\sum x_i y_i}{\sum x_i^۲} = \frac{S_x}{S_y}\beta$$

که β برابر با ضریب X در معادله $Y_i = \alpha + \beta X_i + u_i$ است. از طرف دیگر می‌توان $\hat{\beta}^*$ را با توجه به $S_x^۲ = \frac{\sum x_i^۲}{n}$ به‌صورت زیر نوشت:

$$\hat{\beta}^* = \frac{S_x}{S_y}\frac{\sum x_i y_i}{n S_x^۲} = \frac{\sum x_i y_i / n}{S_x S_y} = r_{xy} \qquad (۲-۴۸)$$

بنابراین $\hat{\beta}^*$ برابر با ضریب همبستگی X و Y است و در نتیجه معادله رگرسیون را برای داده‌های استاندارد شده به‌صورت زیر می‌نویسیم:

$$\hat{Y}_i^* = \alpha^* + \hat{\beta}^* X_i^* = r_{xy} X_i^* \quad ; \quad \hat{\alpha}^* = 0 \,,\, \hat{\beta}^* = r_{xy}$$



نمودار ۱۲-۲: معادله رگرسیون با متغیرهای استانداردشده

برآورد رگرسیون متغیرهای استاندارد شده با Eviews

رگرسیون متغیرهای استاندارد شده را می‌توان با فرمان زیر برآورد نمود:

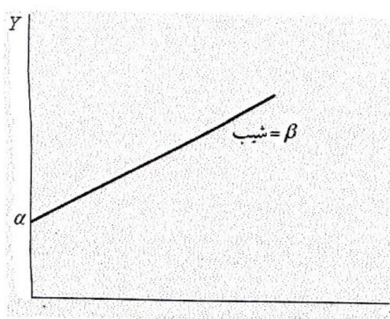
$$\text{LS} \quad \frac{\text{Y}-\text{@mean(Y)}}{\text{@std(Y)}} \quad \frac{\text{X}-\text{@mean(X)}}{\text{@std(X)}}$$

که ضریب $\frac{\text{X}-\text{@mean(X)}}{\text{@std(X)}}$ برابر با ضریب همبستگی بین X و Y است.

۷-۱۵-۲ برآورد معادله روند

گاهی اوقات برخی از متغیرها دارای روند معینی در طول زمان هستند. اگر روند Y خطی باشد، یک معادله خطی را برای آن تعریف می‌کنیم:

$$Y_t = \alpha + \beta t + u_t \qquad (۲-۴۹)$$



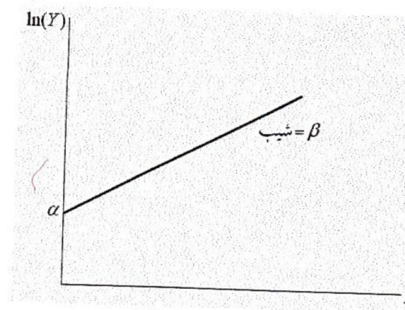
نمودار ۱۳-۲: خط روند

اگر روند غیرخطی باشد می‌توان از معادله زیر استفاده نمود.

$$Y_t = ae^{\beta t}e^{u_t} \quad (۲-۵۰)$$

با گرفتن لگاریتم از طرفین خواهیم داشت:

$$\ln Y_t = \alpha + \beta t + u_t \quad (۲-۵۱)$$



نمودار ۱٤-۲: خط روند لگاریتمی

در معادله (۵۱-۲) شیب خط (β) برابر با نرخ رشد Y است.

برآورد معادله روند با Eviews

برای برآورد روند خطی، فرمان زیر را اجرا می‌کنیم:

```
LS    Y   C   @trend
```

@trend بیانگر متغیر روند t می‌باشد که ضریب آن برابر با β است. برای برآورد معادله روند غیرخطی از فرمان زیر استفاده می‌کنیم:

```
LS    ln(Y)   C   @trend
```

۸-۱۵-۲ برآورد معادله نرخ رشد

برای برآورد معادله نرخ رشد، لازم است که تابع نمایی زیر را در نظر بگیریم.

$$Y_t = ae^{\beta t}e^{u_t} \quad (۲-۵۲)$$

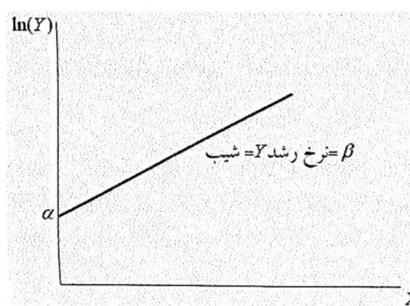
که لگاریتم آن عبارت است از:

$$\ln Y_t = \alpha + \beta t + u_t \ ; \ \alpha = \ln \alpha \quad (۲-۵۳)$$

با مشتق‌گیری نسبت‌به t، نرخ رشد Y (که با $\dot{Y}$ نشان می‌دهیم) به‌دست می‌آید.

$$\dot{Y} = \frac{d\ln Y}{dt} = \beta \quad (۲-۵۴)$$

زیرا $\frac{d\ln Y}{dt}$ برابر با $\frac{dY/t}{Y}$ است که معادل با نرخ رشد Y می‌باشد. بنابراین، ضریب t برابر با متوسط نرخ رشد می‌باشد. شیب خط $\ln Y$ در مقابل t، بیانگر متوسط نرخ رشد Y خواهد بود.



نمودار ۱۵-۲: بر آورد نرخ رشد متوسط

برآورد معادله رشد با Eview

بدین منظور فرمان زیر را اجرا می کنیم:

```
LS    ln(Y)  C  @trend
```

ضریب @trend برابر با β، یعنی متوسط نرخ رشد Y می‌باشد.

مثال ۱۴-۲: بر اساس مثال ۲-۲، معادله زیر را برآورد کنید:

$$Y_i = aX_i^{\beta}$$

ابتدا معادله فوق را به‌صورت لگاریتمی می‌نویسیم:

$$\log(Y_i) = \log(a) + \beta\log(X_i)$$

معادله فوق بر حسب لگاریتم متغیرها به‌صورت خطی است و لذا می‌توان آن را با روش OLS برآورد نمود. نتایج برآورد به‌صورت زیر است:

$$\log(Y_i) = ۱/۴۳۵ + ۰/۵۷۴\log(X_i) \quad R^۲ = ۰/۷۵۶$$

$$(۱۰/۳) \ (۷/۵) \qquad F = ۵۵/۷$$

در این مثال $\beta = ۰/۵۷۴$ و $a = ۱۰^{۱/۴۳۵} = ۲۷/۲۲۷$ می‌باشد. توجه شود که چون $\beta = \frac{\partial\log(Y)}{\partial\log(X)}$ است، لذا β بیانگر درصد تغییر Y به ازای درصد تغییر X است که معروف به کشش Y نسبت‌به X می‌باشد.

برآورد معادلات غیرخطی در Eviews

همان‌طور که اشاره شد برای برآورد معادلات غیر خطی در Eviews می‌توان ابتدا معادله را با تبدیل‌های مناسب، خطی نمود و سپس آن را مشابه معادلات خطی برآورد نمود. به‌عنوان مثال معادله $Y=ae^{\beta X}$ را ابتدا به‌صورت $\log(Y)=\alpha+\beta X$ نوشته که $\alpha=\log(a)$ است، و سپس با فرمان زیر آن را مانند یک معادله خطی برآورد می کنیم:

```
LS    log(Y)  C  X
```

و یا می‌توان آن را با فرمان NLS (حداقل مربعات غیرخطی) به‌صورت زیر برآورد نمود:

NLS   Y=C(1)*exp(C(2)*X)

روش دیگر این است که از منوی Quick گزینه Estimate Equation را انتخاب کرده که به‌دنبال آن پنجره Equation Specification باز می‌شود. در این پنجره، معادله مورد نظر را یا به‌صورت خطی با فرمان log(Y) C X وارد می‌کنیم و یا به‌صورت غیرخطی یعنی Y=C(1)*exp(C(2)*X) وارد می‌کنیم و با انتخاب OK نتایج مربوطه به‌دست می‌آید.

۲-۱۶ تحلیل همبستگی

تحلیل همبستگی به بررسی رابطه بین دو یا چند متغیر می‌پردازد که معمولاً با کوواریانس یا ضریب همبستگی بیان می‌شود.

۲-۱۶-۱ کوواریانس

برای بیان همبستگی خطی بین دو متغیر X و Y، یکی از معیارهای مهم و کاربردی، کوواریانس می‌باشد که به‌صورت زیر تعریف می‌گردد:

$$\operatorname{cov}(X,Y)=\sigma_{xy}=E(X-\mu_x)(Y-\mu_y)=E(XY)-\mu_x\mu_y \qquad (۲-۲۶)$$

μ_x و μ_y به ترتیب امید ریاضی X و Y هستند. کوواریانس X و Y را با $\operatorname{cov}(X,Y)$ یا σ_{xy} و تخمین آن را بر حسب داده‌های نمونه با $\hat{\sigma}_{xy}$ نشان می‌دهیم که برابر است با:

$$\hat{\sigma}_{xy}=\frac{\sum_{i=1}^{n}x_i y_i}{n}=\frac{\sum_{i=1}^{n}x_i Y_i}{n}=\frac{\sum_{i=1}^{n}X_i y_i}{n}$$

$$=\frac{\sum X_i Y_i-\frac{(\sum X_i)(\sum X_i)}{n}}{n}=\frac{\sum X_i Y_i}{n}-\overline{X}\overline{Y}$$

مشکلی که در رابطه با کوواریانس وجود دارد این است که شدت و ضعف همبستگی خطی بین X و Y را بیان نمی‌کند. به‌عبارت دقیق‌تر مقدار آن کاملاً تحت تأثیر بزرگی و کوچکی مقادیر متغیرها قرار دارد. بدین منظور به مثال‌های ۲-۱۵ و ۲-۱۶ توجه کنید.

مثال ۱۵-۲: برای بررسی رابطه بین دو متغیر X و Y نمونه‌ای به حجم $n = ۵$ انتخاب شد. که برای X مقادیر ۱، ۳، ۲، ۴، ۸ و برای Y نیز مقادیر ۲، ۲، ۴، ۶ و ۶ به‌دست آمده است. تخمین کوواریانس X و Y عبارت است از:

X_i	Y_i	X_iY_i
۱	۲	۲
۳	۲	۶
۲	۴	۸
۴	۶	۲۴
۸	۶	۴۸
۱۸	۲۰	۸۸

$$\overline{X} = \frac{۱۸}{۵} = ۳/۶ \quad , \quad \overline{Y} = \frac{۲۰}{۵} = ۴$$

$$\operatorname{cov}(X,Y) = \hat{\sigma}_{xy} = \frac{۸۸}{۵} - (۳/۶)(۴) = ۳/۲$$

چون $\hat{\sigma}_{xy}$ مثبت است پس بین X و Y رابطه مثبت وجود دارد.

مثال ۱۶-۲: فرض کنید که مشاهدات X و Y در مثال ۱۵-۲ در ۱۰۰ ضرب شده باشند. کوواریانس X و Y را حساب کنید.

X_i	Y_i	X_iY_i
۱۰۰	۲۰۰	۲۰۰۰۰
۳۰۰	۲۰۰	۶۰۰۰۰
۲۰۰	۴۰۰	۸۰۰۰۰
۴۰۰	۶۰۰	۲۴۰۰۰۰
۸۰۰	۶۰۰	۴۸۰۰۰۰
۱۸۰۰	۲۰۰۰	۸۸۰۰۰۰

$$\overline{X} = \frac{۱۸۰۰}{۵} = ۳۶۰ \quad , \quad \overline{Y} = \frac{۲۰۰۰}{۵} = ۴۰۰$$

$$\hat{\sigma}_{XY} = \frac{۸۸۰۰۰۰}{۵} - (۳۶۰)(۴۰۰) = ۳۲۰۰۰$$

در اینجا نیز چون کوواریانس مثبت است پس بین X و Y رابطه مثبت وجود دارد. در این مثال، کوواریانس X و Y نسبت‌به مثال قبلی ۱۰۰۰۰ برابر شده است که ناشی از بزرگ شدن مقیاس متغیرها است. اما بدیهی است که شدت همبستگی X و Y در هر دو حالت، یکسان است.

خواص کوواریانس

۱- اگر به X یا Y و یا به هر دو، مقدار ثابت a را اضافه یا کم کنیم، کوواریانس تغییری نمی‌کند.

$$\operatorname{cov}(X+a,Y)=\frac{\sum[(X_i+a)-(\bar{X}+a)](Y_i-\bar{Y})}{n}$$
$$=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n}=\operatorname{cov}(X,Y)$$

و در حالت کلی خواهیم داشت:

$$\operatorname{cov}(X\pm a,Y\pm b)=\operatorname{cov}(X,Y)$$

۲- اگر مقدار Xها یا Yها را در عدد ثابت a ضرب کنیم، کوواریانس نیز در a ضرب می‌شود.

$$\operatorname{cov}(aX,Y)=\frac{\sum(aX_i-a\bar{X})(Y_i-\bar{Y})}{n}$$
$$=\frac{a\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n}=a\operatorname{cov}(X,Y)$$

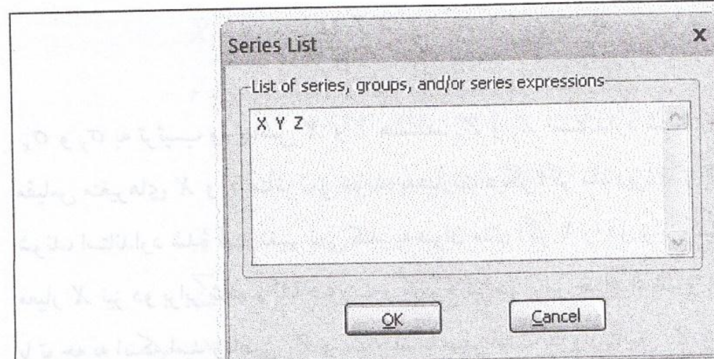
$$\operatorname{cov}(X,aY)=\frac{\sum(X_i-\bar{X})(aY_i-a\bar{Y})}{n}$$
$$=\frac{a\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n}=a\operatorname{cov}(X,Y)$$

۳- اگر Xها در a و Yها در b ضرب شوند، کوواریانس در ab ضرب خواهد شد.

$$\operatorname{cov}(aX,bY)=\frac{\sum(aX_i-a\bar{X})(bY_i-b\bar{Y})}{n}$$
$$=\frac{ab\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n}=ab\operatorname{cov}(X,Y)$$

محاسبه کوواریانس با Eviews

برای محاسبه کوواریانس با Eviews می‌توان در سطر فرمان عبارت cov(X,Y) را اجرا نمود. همچنین می‌توان از منوی Quick گزینه Group Statistics را انتخاب نمود که در این صورت پنجره Series List به‌صورت زیر باز می‌شود که نام متغیرهای مورد نظر را در آن وارد می‌کنیم:



نتایج محاسبهٔ کوواریانس‌ها به‌صورت زیر نشان داده می‌شود:

Group: UNTITLED Workfile: INVESTMENT::Unt...

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Covariance

	X	Y	Z
X	546006.3	28695.27	126031.3
Y	28695.27	2056.391	7316.231
Z	126031.3	7316.231	31671.92

قطر اصلی، واریانس هر یک از متغیرها را نشان می‌دهد، به‌عنوان مثال واریانس X برابر با ۵۴۶۰۰۶/۳ می‌باشد. سایر عناصر، کوواریانس بین متغیرها را نشان می‌دهد، مثلاً کوواریانس بین X و Y برابر با ۲۸۶۹۵/۲۷ و کوواریانس بین X و Z برابر با ۱۲۶۰۳۱/۳ و کوواریانس بین Y و Z برابر با ۷۳۱۶/۲۳۱ می‌باشد.

۲-۱۶-۲ ضریب همبستگی

از مثال‌های ۱۵-۲ و ۱۶-۲ روشن است که کوواریانس کاملاً متأثر از مقدار متغیرها است. به‌عنوان نمونه، مثال ۱۵-۲ بیانگر اندازه‌گیری X و Y برحسب متر است ولی در مثال ۱۶-۲ مقادیر X و Y بر حسب سانتی‌متر اندازه‌گیری شده است. بدیهی است که با تغییر مقیاس، شدت و ضعف همبستگی بین X و Y دچار تغییر نخواهد شد، در حالی که کوواریانس کوچک و بزرگ می‌شود.

برای حل این مشکل باید ابتدا متغیرهای X و Y را استاندارد نمود تا مقیاس اندازه‌گیری متغیرها تأثیر خود را از دست بدهد و سپس کوواریانس آنها را محاسبه کنیم:

$$Z_x = \frac{X - \mu_X}{\sigma_x} \qquad , \quad Z_y = \frac{Y - \mu_y}{\sigma_y}$$

σ_x و σ_y به ترتیب واریانس X و Y هستند. Z_x و Z_y استاندارد شدهٔ X و Y می‌باشند و از مقیاس متغیرهای X و Y متأثر نمی‌شوند. به‌عبارت دیگر اگر مقادیر X و Y بزرگ یا کوچک شوند، استاندارد شدهٔ آنها تغییر نمی‌کند. به‌عنوان مثال اگر X را دو برابر کنیم، میانگین و انحراف معیار X نیز دو برابر شده و لذا صورت و مخرج نیز دو برابر خواهند شد و Z_x ثابت می‌ماند. حال با توجه به اینکه امید ریاضی Z_x و Z_y برابر با صفر است کوواریانس Z_x و Z_y را که موسوم به ضریب همبستگی X و Y است به‌صورت زیر محاسبه می‌کنیم:

$$\begin{aligned} \operatorname{cov}(Z_x, Z_y) = \sigma_{Z_x, Z_y} &= E(Z_x Z_y) = E\left(\frac{X - \mu_x}{\sigma_x}\right)\left(\frac{Y - \mu_y}{\sigma_y}\right) \\ &= \frac{E(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \rho_{xy} \end{aligned} \qquad (۲-۳۴)$$

تخمین ضریب همبستگی بر اساس داده‌های نمونه را با $\hat{\rho}_{xy}$ یا r_{xy} و یا به‌طور خلاصه با $\hat{\rho}$ یا r نشان می‌دهیم:

$$\begin{aligned} \hat{\rho} = r &= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_y^2}} = \frac{\frac{\sum x_i y_i}{n}}{\sqrt{\frac{\sum x_i^2}{n} \times \frac{\sum y_i^2}{n}}} \\ &= \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)\left(\sum Y_i^2 - n\bar{Y}^2\right)}} \end{aligned} \qquad (۲-۳۵)$$

به‌سادگی می‌توان ثابت نمود که اگر X و Y در مقدار ثابتی مانند a ضرب یا تقسیم شوند مقدار r هیچ تغییری نخواهد کرد.

مثال ۱۷-۲: در مثال ۱۵-۲و ۱۶-۲ ضریب همبستگی X و Y را محاسبه کنید.

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
۱	۲	۱	۴	۲
۳	۲	۹	۴	۶
۲	۴	۴	۱۶	۸
۴	۶	۱۶	۳۶	۲۴
۸	۶	۶۴	۳۶	۴۸
۱۸	۲۰	۹۴	۹۶	۸۸

$$\overline{X}=۳/۶\ ,\ \overline{Y}=۴\ ,\ \sum X_i^۲=۹۴\ ,\ \sum Y_i^۲=۹۶\ ,\ \sum X_iY_i=۸۸$$

$$r=\frac{۸۸-(۵)(۳/۶)(۴)}{\sqrt{(۹۴-۵\times ۳/۶^۲)(۹۶-۵\times ۴^۲)}}=\frac{۱۶}{\sqrt{(۲۹/۲)(۱۶)}}=۰/۷۴۰$$

این محاسبات را اگر برای مثال ۱۶-۲ انجام دهیم، خواهیم داشت:

$$\overline{X}=۳۶۰\quad ,\ \overline{Y}=۴۰۰$$

$$\sum X_i^۲=۹۴۰۰۰۰\quad ,\ \sum Y_i^۲=۹۶۰۰۰۰\quad ,\ \sum X_iY_i=۸۸۰۰۰۰$$

با قراردادن این مقادیر در فرمول r، مقدار ضریب همبستگی همچنان برابر با ۰/۷۴ خواهد شد.

خواص ضریب همبستگی

۱- اگر به X ها یا Y ها و یا هر دو آنها مقدار ثابتی را اضافه یا کم کنیم، ضریب همبستگی تغییری نخواهد کرد.

$$r=\frac{\sum[(X_i+a)-(\overline{X}+a)][(Y_i+b)-(\overline{Y}+b)]}{\sqrt{\sum[(X_i+a)-(\overline{X}+a)]^۲\sum[(Y_i+b)-(\overline{Y}+b)]^۲}}$$

$$=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^۲\sum(Y_i-\overline{Y})^۲}}$$

۲- اگر X ها یا Y ها و یا هر دو را در عدد ثابتی ضرب یا تقسیم کنیم، ضرب همبستگی تغییری نخواهد کرد.

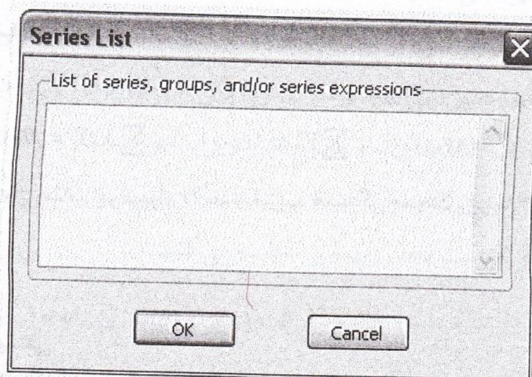
$$r_{ax,by}=\frac{\sum(aX_i-a\overline{X})(bY_i-b\overline{Y})}{\sqrt{\sum(aX_i-a\overline{X})^۲\sum(bY_i-b\overline{Y})^۲}}$$

$$=\frac{ab\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{a^۲\sum(X_i-\overline{X})^۲b^۲\sum(Y_i-\overline{Y})^۲}}$$

$$=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^۲\sum(Y_i-\overline{Y})^۲}}=r$$

محاسبه ضریب همبستگی با Eviews

برای محاسبه ضریب همبستگی با Eviews می‌توان از منوی Quick گزینه Group Statistics و سپس Correlations را انتخاب که در این صورت پنجره Series List باز می‌شود.



پس از ورود نام متغیرهای مورد نظر و انتخاب OK، نتایج محاسبهٔ ضرایب همبستگی به‌صورت زیر نشان داده می‌شود:

Group: UNTITLED Workfile: INVESTMENT::Untitled\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Correlation

	X	Y	Z
X	1.000000	0.856364	0.958390
Y	0.856364	1.000000	0.906562
Z	0.958390	0.906562	1.000000

قطر اصلی برابر ۱ است زیرا ضریب همبستگی هر متغیر با خودش برابر با ۱ خواهد بود. سایر عناصر ضریب همبستگی بین متغیرها را نشان می‌دهد، مثلاً ضریب همبستگی بین X و Y برابر با ۰/۸۵۶ و ضریب همبستگی بین X و Z برابر با ۰/۹۵۸ و ضریب همبستگی بین Y و Z برابر با ۰/۹۰۶ می‌باشد.

آزمون فرضیه راجع به ضریب همبستگی

ضریب همبستگی خطی بین X و Y بیانگر شدت و ضعف رابطهٔ خطی بین X و Y است. مقدار r بین ۱+ و ۱- می‌باشد. هر چه r به صفر نزدیک باشد، رابطه خطی بین X و Y ضعیف است و هر چه که به ۱+ و ۱- نزدیک باشد گفته می‌شود که X و Y رابطه قوی با هم دارند که در حالت اول رابطه قوی مثبت و در حالت دوم رابطه قوی دارند. اما سؤال این است r در چه محدوده‌ای باید باشد تا بگوییم که X و Y با هم رابطه‌ای ندارند. برای قضاوت در این مورد،

مجدداً به آزمون فرضیه متوسل می‌شویم. برای انجام آزمون $\rho=0$ در مقابل $\rho\neq 0$ از آزمون t استفاده می‌شود:[1]

$$t_{n-2}=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \qquad (2\text{-}36)$$

اگر مقدار t از مقدار بحرانی (که معمولاً برابر با ۲ می‌باشد) بزرگتر باشد، فرضیه $\rho=0$ رد می‌شود و ضریب همبستگی از نظر آماری معنادار است.

مثال ۱۸-۲: در مثال ۲-۲ ضریب همبستگی برابر با ۰/۸۵۹ می‌باشد. آیا ضریب همبستگی از نظر آماری معنادار است؟

$$t_{n-2}=\sqrt{20-2}\frac{0/859}{\sqrt{1-0/859^2}}=7/118$$

چون مقدار $t=7/118$ در ناحیه بحرانی قرار می‌گیرد، لذا فرضه $\rho=0$ رد می‌شود و ضریب همبستگی بین X و Y تفاوت معناداری از صفر دارد.

مثال ۱۹-۲: در مثال ۱۷-۲ ضریب همبستگی برابر با ۰/۷۴ می‌باشد. آیا ضریب همبستگی از نظر آماری معنادار است؟

$$t_{n-2}=\sqrt{5-2}\frac{0/74}{\sqrt{1-0/74^2}}=1/91$$

چون مقدار $t=1/91$ در ناحیه بحرانی قرار نمی‌گیرد (در اینجا عدد بحرانی برابر با $t_{0/025,3}=3/182$ است)، لذا فرضه $\rho=0$ رد نمی‌شود و ضریب همبستگی بین X و Y تفاوت معناداری از صفر ندارد.

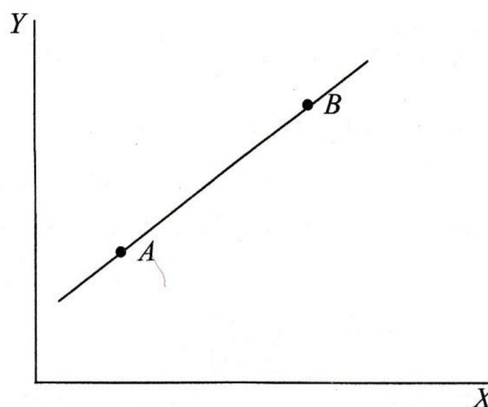
در مثال ۱۹-۲ علی‌رغم اینکه ضریب همبستگی برابر با ۰/۷۴ است، ولی به‌دلیل اندک بودن تعداد مشاهدات، تفاوت معناداری با صفر ندارد. این در حالی است که ضریب همبستگی ۰/۲ با ۲۰۰ مشاهده و یا ضریب همبستگی ۰/۴ با ۳۰ مشاهده، از نظر آماری معنادار است. توجه شود که اگر فقط دو مشاهده داشته

---

[1] رابطه بین r و توزیع t با درجه آزادی $n-2$ دقیقاً معادل با توزیع t است که در آزمون معنادار بودن ضریب β به‌کار می‌رود. اگر در t به جای $\hat{\beta}$ و $\hat{\sigma}^2$ قرار داده و نتیجه را ساده کنیم خواهیم داشت:

$$t_{n-2}=\frac{\hat{\beta}}{\sqrt{\text{var}(\hat{\beta})}}=\sqrt{\sum x_i^2}\frac{\hat{\beta}}{\sqrt{\hat{\sigma}^2}}=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

باشیم در این صورت ضریب همبستگی برابر با ۱ است و خط رگرسیون دقیقاً از این دو نقطه می‌گذرد، ولی از نظر آماری معنادار نیست.



نمودار ۱٦-۲: همبستگی خطی با دو مشاهده

مسائل

۱- با توجه به اطلاعات زیر مقادیر $\hat{\alpha}$، $\hat{\beta}$ و $SE(\hat{\beta})$ را حساب نمایید؟ آیا $\hat{\beta}$ از لحاظ آماری معنادار است؟

$$n=۴۰ \ , \ \overline{X}=۴۲۰ \ , \ \overline{Y}=۱۰۰ \ , \ \sum X_iY_i=۸۵۰۰۰۰$$
$$\sum X_i^۲=۴۰۰۰۰۰۰ \ , \ RSS=۱۴۰$$

۲- مدل‌های زیر مفروض هستند:

$$\ln Y_i=\beta_۱+\beta_۲\ln X_i+u_i$$
$$\ln Y_i^*=\alpha_۱+\alpha_۲\ln X_i^*+u_i^*$$

که $Y_i^*=w_۱Y_i$ و $X_i^*=w_۲X_i$ بوده $w_۱$ و $w_۲$ ثابت هستند. آیا $R^۲$ در این دو مدل، متفاوت است؟ ثابت کنید.

۳- در مسئله ۱ فصل اول، رگرسیون Y روی X را برآورد کرده و نتایج آن را تحلیل نمایید.

۴- در مسئله ۱ فصل اول، تغییرات Y را روی تغییرات X برازش کرده و نتایج آن را تحلیل و با مسئله ۳ مقایسه نمایید.

۵- در مسئله ۱ فصل اول، ضریب همبستگی بین X، Y و Z را حساب کرده و آزمون معنی‌دار بودن آنها را انجام دهید.

۶- در مسئله ۱ فصل اول، ضریب همبستگی بین نرخ رشد متغیرهای X، Y و Z را حساب کرده و آزمون معنی‌دار بودن آنها را انجام دهید. نتایج را با مسئله ۵ مقایسه کنید.

۷- مقدار آماره t برای ضریب X در مسئله ۳ را با مقدار آماره t در آزمون معنی‌دار بودن ضریب همبستگی بین X و Y مقایسه کنید. دلیل تفاوت یا مشابهت چیست؟

۸- ثابت کنید که $\sum X_i e_i = 0$ است.

۹- ثابت کنید که $\sum \hat{Y}_i e_i = 0$ است.

۱۰- معادله رگرسیون برای $n = 30$ مشاهده به‌صورت زیر برآورد شده است:

$$\hat{Y}_i = 2/5 + 4/2X_i \quad , \quad R^2 = 0/8$$

مقدار آماره t برای ضریب X_i و آماره F را حساب کنید.

۱۱- در مسئله ۱۰، فرضیه $\beta = 5$ را آزمون کنید.

۱۲- در معادله $Y_i = \alpha + \beta X_i + u_i$ مقادیر متغیرها بر حسب هزار ریال بیان شده است. اگر مقادیر متغیرها را بر حسب ریال بیان کنیم آیا مقدار تخمینی ضرایب، تغییر خواهد کرد. چرا؟

۱۳- در مسئله ۳، رگرسیون Y روی X را برای دوره ۷۰-۱۳۵۱ برآورد کرده و سپس مقادیر Y را برای دوره ۸۰-۱۳۷۱ پیش‌بینی کرده و با مقادیر واقعی مقایسه کنید.

۱۴- مفهوم فرض $E(u_i \mid X_i)$ را به‌طور مستدل توضیح دهید.

۱۵- مقادیر زیر را داریم:

$$\sum X_i = 130 \quad , \quad \sum Y_i = 240 \quad , \quad n = 20$$
$$\sum X_i^2 = 1000 \quad , \quad \sum Y_i^2 = 3500 \quad , \quad \sum X_i Y_i = 1800$$

معادله $Y_i = \alpha + \beta X_i + u_i$ را برآورد کرده و مقادیر t، F و R^2 را حساب کنید.

۱۶- در مسئله ۱۵، ضریب هبستگی X و Y را حساب کرده و آزمون معنادار بودن آن را انجام دهید.

۱۷- ثابت کنید که $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum x_i^2})$ است.

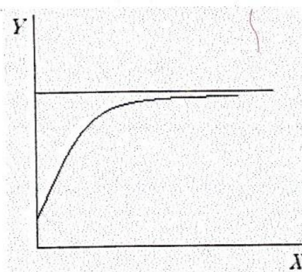
۱۸- ثابت کنید که $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$ است.

۱۹- در رگرسیون $Y_i = \alpha + \beta D_i + u_i$ متغیر D_i فقط مقادیر صفر و یک را اختیار می‌کند. مقدار D_i برای n_1 مشاهده برابر با ۱ و برای بقیه مشاهدات (n_2) برابر با صفر است ($n_1 + n_2 = n$). تخمین‌زنندهٔ α و β را به‌دست آورده و واریانس آنها را حساب کنید.

۲۰- در رگرسیون $Y_i = \alpha + \beta X_i + u_i$ امید ریاضی و واریانس شرطی را حساب کنید.

۲۱- ثابت کنید که اگر دو متغیر X و Y مستقل باشند، در آن صورت امید ریاضی شرطی Y مستقل از X خواهد بود.

۲۲-اگر مشاهدات X و Y به‌صورت نمودار زیر باشد، چه معادله‌ای را برای آن پیشنهاد می‌کنید.



۲۲- در معادله $Y_t = \alpha + \beta X_t + u_t$ کوواریانس $\hat{\alpha}$ و $\hat{\beta}$ را حساب کنید. آیا امکان دارد که $\text{cov}(\hat{\alpha}, \hat{\beta}) = 0$ باشد.

۲۳- ثابت کنید که $\sum \hat{y}_t^2 = \sum \hat{y}_t y_t$ است.

۲۴- ثابت کنید که $\Sigma \hat{y} e_t = 0$ است.

۲۵- ثابت کنید که معادله $\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$ از نقطه $(\overline{X}, \overline{Y})$ می‌گذرد. نمودار آن را ترسیم کنید.

۲۶- ثابت کنید که $\Sigma e_t = 0$ معادل با $\overline{Y} = \hat{\alpha} + \hat{\beta} \overline{X}$ است.

۲۷- در معادله $Y_t = \alpha + \beta X_t + u_t$ اگر u_t و X_t مستقل نباشند، واریانس $\hat{\beta}$ را حساب کنید. آیا با عبارت $\frac{\sigma^2}{\Sigma x_t^2}$، متفاوت خواهد بود یا نه؟

۲۸- فرض کنید که در معادله $Y_t = \alpha + \beta X_t + u_t$ ، متغیرهای u_t و X_t همبستگی مثبت داشته باشند. اگر $\beta > ۰$ باشد، آنگاه ثابت کنید که تخمین β از روش OLS منجر به برآورد بیش از حد β خواهد شد.

۲۹- اگر X_t و u_t در معادله $Y_t = \alpha + \beta X_t + u_t$ مستقل نباشند، آیا تخمین α از روش OLS دچار تورش خواهد شد؟

۳۰- ثابت کنید که در رگرسیون $Y_t = \alpha + \beta X_t + u_t$، تغییرات توضیح داده شده برابر است با:

$$ESS = \hat{\beta}^2 \sum x_t^2 = \hat{\beta} \sum x_t y_t$$

۳۱- اگر معادله $Y_t = a + ae^{\beta X_t}$ را داشته باشیم، اولاً نمودار آن را رسم کرده و ثانیاً چگونه آن را تبدیل به یک مدل قابل تخمین و خطی می‌کنید.

۳۲- اگر رگرسیون فقط دارای عرض از مبدأ باشد ($Y_i = \alpha + u_i$)، ثابت کنید که $\hat{\alpha} = \overline{Y}$ و $\text{var}(\hat{\alpha}) = \frac{\sigma^2}{n}$ می‌باشد.

۳۳- رگرسیون‌های زیر را داریم:

$$۱)\ Y_t = \alpha + \beta X_t + u_t$$

$$۲)\ X_t = a + bY_t + v_t$$

ضریب تعیین معادله (۱) و (۲) را به‌ترتیب با R_1^2 و R_2^2 نشان می‌دهیم.

الف) ثابت کنید که $R_1^2 = R_2^2$ است.

ب) ثابت کنید که $\hat{\beta} \leq \frac{1}{\hat{b}}$ است.

ج) آیا تساوی $\sum \hat{u}_t^2 = \sum \hat{v}_t^2$ برقرار است؟

۳۴- رگرسیون‌های زیر برآورد شده‌اند:

$$Y_t = ۲/۴۵ + ۱/۲۵ X_t$$
$$X_t = ۴.۵۲ + ۰/۸۴ Y_t$$

آیا نتایج فوق درست است؟

۳۵- معادلات زیر برای درآمد و مصرف برآورد شده‌اند:

$$Y_t = \alpha + ۱/۲۵ C_t$$
$$C_t = a + ۰/۷۵ Y_t$$

الف) ضریب تعیین (R^2) را برای هر یک از این معادلات حساب کنید.

ب) آیا ضرایب تخمینی، معنادار هستند؟

ج) اگر $Y=C+S$ و S پس‌انداز باشد، ضریب همبستگی بین Y و S و بین C و S را حساب کنید.

۳۶- مقادیر Y_t روی زمان (t) برازش می‌شود که t از سال ۱۳۶۱ تا ۱۳۹۱ می‌باشد.

$$Y_t=\alpha+\beta t+u_t$$

ابتدا متغیر زمان را با مقادیر $t=۱۳۶۱,۱۳۶۲,...,۱۳۹۰$ نشان داده و α و β را برآورد می‌کنیم که با $\hat{\alpha}_1$ و $\hat{\beta}_1$ نشان می‌دهیم. حال برای متغیر زمان مقادیر $t=۱,۲,...,۳۰$ را تعریف کرده و مجدداً معادله را برآورد کرده و با $\hat{\alpha}_2$ و $\hat{\beta}_2$ نشان می‌دهیم. چه تفاوتی بین $\hat{\alpha}_1$ با $\hat{\alpha}_2$ و همچنین $\hat{\beta}_1$ با $\hat{\beta}_2$ وجود دارد؟

۳۷- معادله $Y_t=aX_t^{\beta}$ را داریم. برای خطی کردن آن از طرفین لگاریتم می‌گیریم:

$$\log Y_t=\alpha+\beta X_t$$
$$\ln Y_t=\alpha+\beta X_t$$

آیا $\hat{\alpha}$ و $\hat{\beta}$ در این دو مدل متفاوتند؟

۳۸- در مدل $Y_t=aX_t^{\beta}$ متغیر تصادفی u_t توزیع نرمال با میانگین صفر و واریانس $\sigma^۲$ دارد. اگر برای برآورد این مدل از شکل لگاریتمی آن استفاده کنیم، خواهیم داشت:

$$\ln Y_t=\alpha+\beta X_t+\ln u_t \quad , \quad \alpha=\ln a$$

الف) توزیع $\ln u_t$ چگونه خواهد بود؟

ب) اگر بخواهیم شرط $E(\ln u_t)=۰$ برقرار باشد، توزیع u_t چگونه خواهد بود؟

۳۹- رگرسیون $\hat{Y}_t=-۲/۵+۰/۸X_t$ با $n=۴۲$ مشاهد برآورد شده است. اگر $R^۲=۰/۴۵$ باشد، فرضیه‌های زیر را آزمون کنید:

الف) $H_۰:\beta=۰$

ب) $H_۰:R^۲=۰$

۴۰- در مسئله ۳۹ اگر $n=۱۰$ باشد فرضیه‌های مورد نظر را آزمون کنید. نتیجه آزمون چه تفاوتی کرد؟ چرا؟

۴۱- در رگرسیون $Y_t=\alpha+\beta X_t+u_t$ وقتی از روش OLS استفاده می‌شود، ثابت کنید که روابط زیر برقرار است:

الف) $\sum e_i = 0$

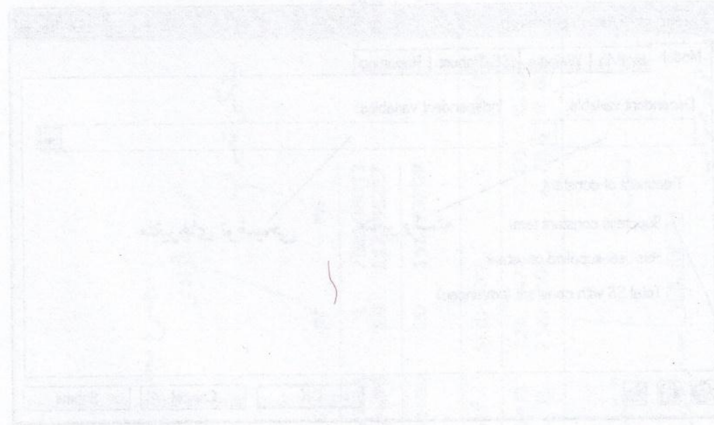
ب) $\sum e_i X_i = 0$

ج) $\sum e_i \hat{Y}_i = 0$

د) $\sum Y_i = \sum \hat{Y}_i$

۴۲- نشان دهید که در رگرسیون $Y_t = \alpha + \beta X_t + u_t$ شرط $\overline{Y} = \overline{\hat{Y}}$ برقرار است ولی در رگرسیون $Y_t = \beta X_t + u_t$، برقرار نیست. عدم برقراری این شرط به چه معنا است؟

۴۳- ثابت کنید که R^2 برابر با ضریب همبستگی $\hat{Y}_t$ و Y_t است.



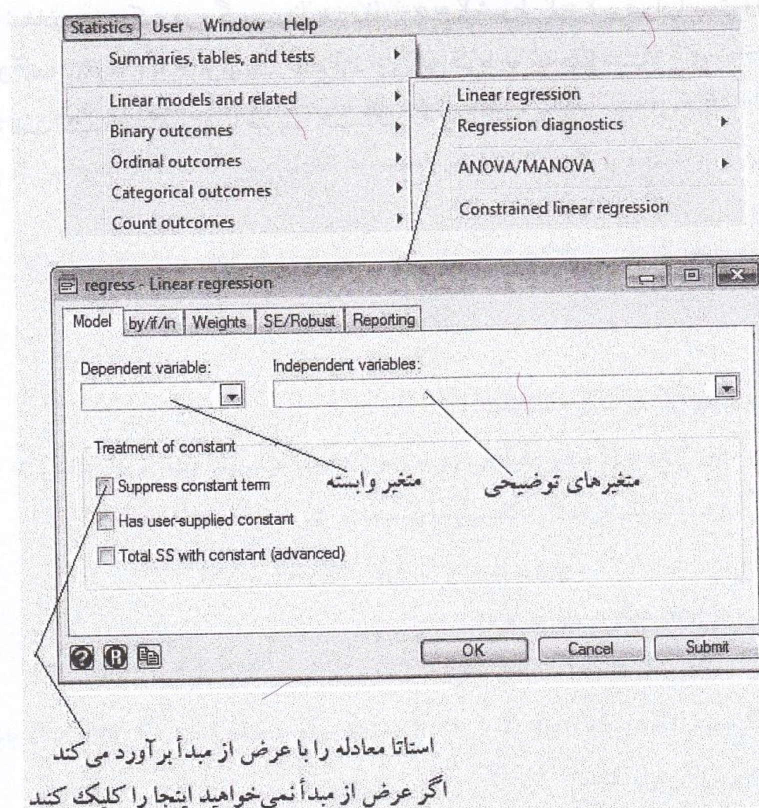
ضمیمه فصل ۲: رگرسیون ساده در Stata

برآورد رگرسیون ساده با Stata

فایل data1

به‌منظور برآورد رگرسیون خطی، مسیر زیر را دنبال می‌کنیم:

Linear regression ← Linear models and related ← Statistics



علاوه بر این، می‌توان معادله را با فرمان زیر تخمین زد:

```
reg y x
```

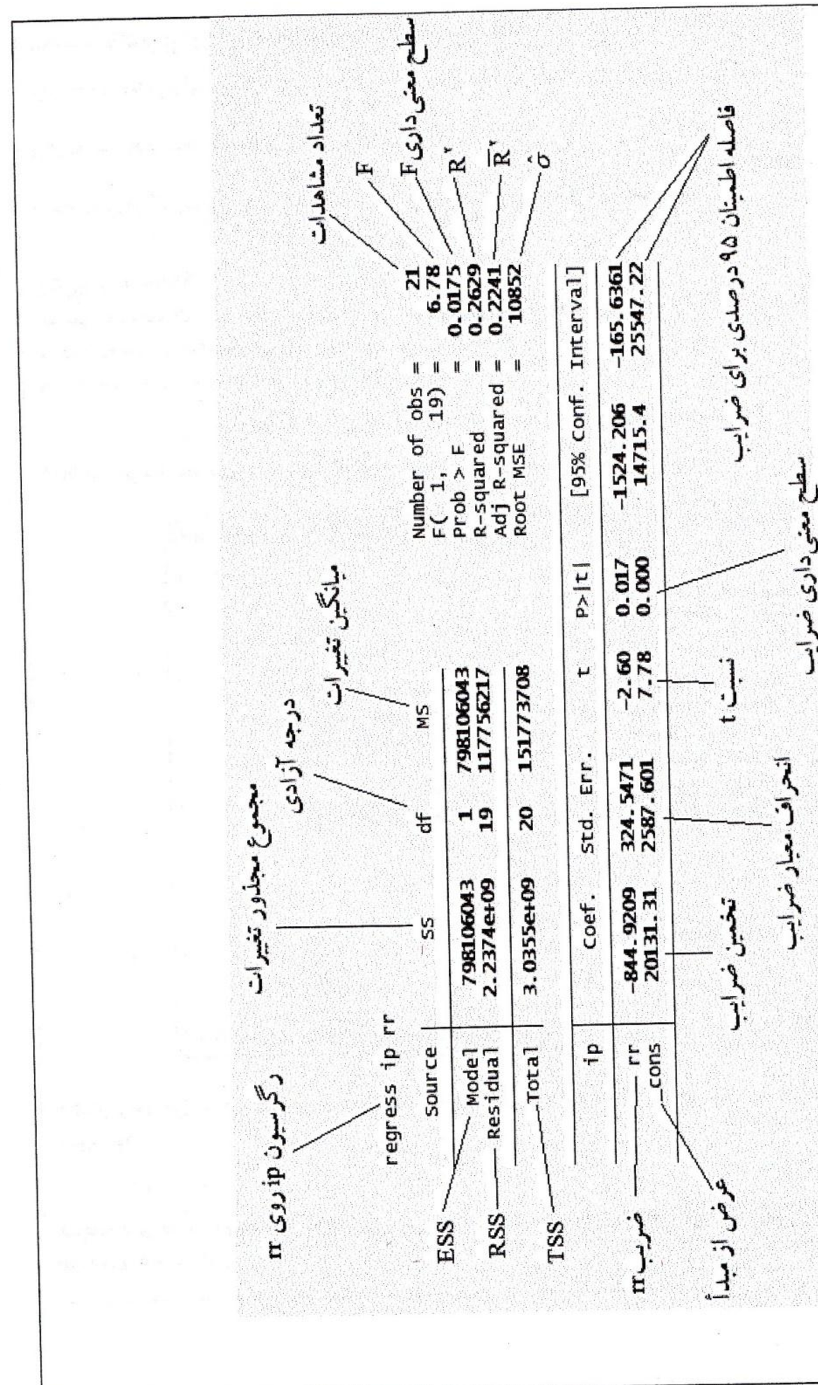
اگر بخواهیم Y را فقط روی عرض از مبدأ رگرس کنیم، از فرمان زیر استفاده می‌کنیم:

```
reg y
```

و برای رگرسیون بدون عرض از مبدأ، فرمان زیر را اجرا می‌کنیم:

```
reg y x, noc
```

بعد از اجرای هر یک از فرمان‌های فوق، نتایج تخمین عبارت است از:



```
regress ip rr

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =    6.78
       Model |   798106043     1   798106043           Prob > F      =  0.0175
    Residual |  2.2374e+09    19   117756217           R-squared     =  0.2629
-------------+------------------------------           Adj R-squared =  0.2241
       Total |  3.0355e+09    20   151773708           Root MSE      =   10852

------------------------------------------------------------------------------
          ip |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |  -844.9209   324.5471    -2.60   0.017    -1524.206   -165.6361
       _cons |   20131.31   2587.601     7.78   0.000      14715.4    25547.22
------------------------------------------------------------------------------
```

محاسبه مقادیر برآوردی و باقیمانده‌ها

برای محاسبهٔ مقادیر برآوردی Y یعنی $\hat{Y}$، بعد از برآورد معادله مورد نظر، فرمان زیر را اجرا می‌کنیم:

```
predict yhat
```

برای محاسبه باقیمانده‌ها (خطاها) نیز فرمان زیر را اجرا می‌کنیم:

```
predict e, residuals
```

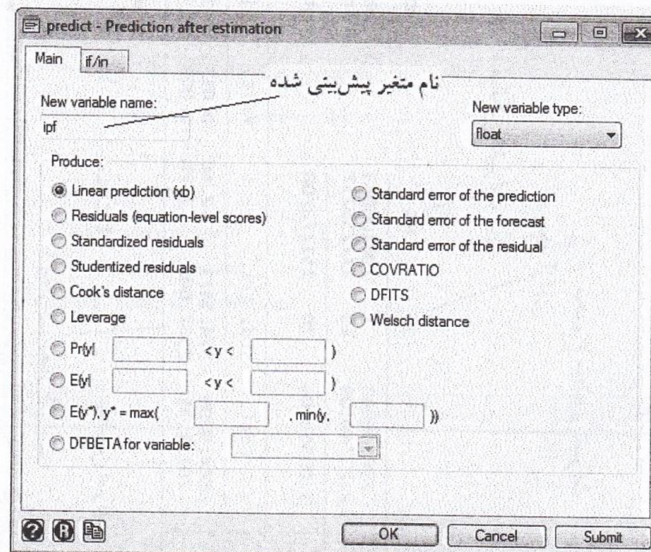
به‌جای روش فوق می‌توان به روشی که در قسمت بعدی، برای پیش‌بینی گفته می‌شود، عمل نمود.

پیش‌بینی با Stata

پیش‌بینی به دو صورت است: درون نمونه‌ای و برون نمونه‌ای. پیش‌بینی درون‌نمونه‌ای مشابه برآورد مقادیر تخمینی است که در بالا به آن اشاره شد. این نوع پیش‌بینی را به صورت زیر نیز می‌توان انجام داد.
بدین منظور، بعد از تخمین مدل، مسیر زیر را دنبال می‌کنیم:

Predictions, Residuals, etc → Postestimation → Statistics

با انتخاب مسیر فوق پنجره زیر باز می‌شود که گزینه‌های مربوط به آن عبارتند از:



با انجام مراحل فوق، مقادیر پیش‌بینی شده ip با نام ipf محاسبه می‌شود. برای مقایسه آنها می توان با فرمان زیر نمودار مربوطه را ترسیم نمود.

```
twoway (line ip ipf t)
```

همچنین برای محاسبهٔ باقیمانده‌ها (خطاهای پیش‌بینی)، مشابه محاسبهٔ مقادیر پیش‌بینی عمل کرده و در پنجرهٔ مربوط به پیش‌بینی، به صورت زیر عمل می‌کنیم:

e را برای باقیمانده‌ها انتخاب می‌کنیم

این گزینه را انتخاب می‌کنیم

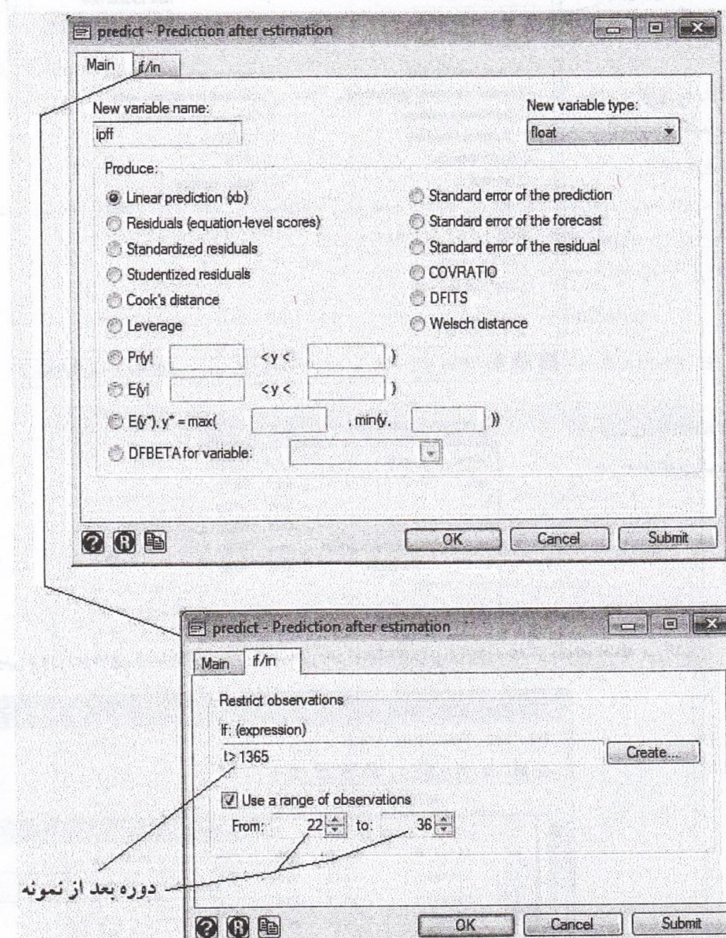
در قسمت variables نام متغیر e اضافه می‌شود

Name	Label	Type	Format
date		int	%8.0g
ip		float	%9.0g
rr		float	%9.0g
ipf	Linear prediction	float	%9.0g
e	Residuals	float	%9.0g

برای پیش‌بینی برون نمونه‌ای، ابتدا مقادیر متغیر توضیحی (در اینجا rr) را برای دوره بعد از نمونه، اضافه می‌کنیم:

مقادیر rr را برای سال‌های بعد از نمونه برابر با ۱ فرض می‌کنیم

حال پنجره predict را باز کرده و نام متغیر موردنظر را (در اینجا ipff) وارد می‌کنیم. سپس منوی if/in را انتخاب کرده و دوره پیش‌بینی برون نمونه‌ای را وارد می‌کنیم.

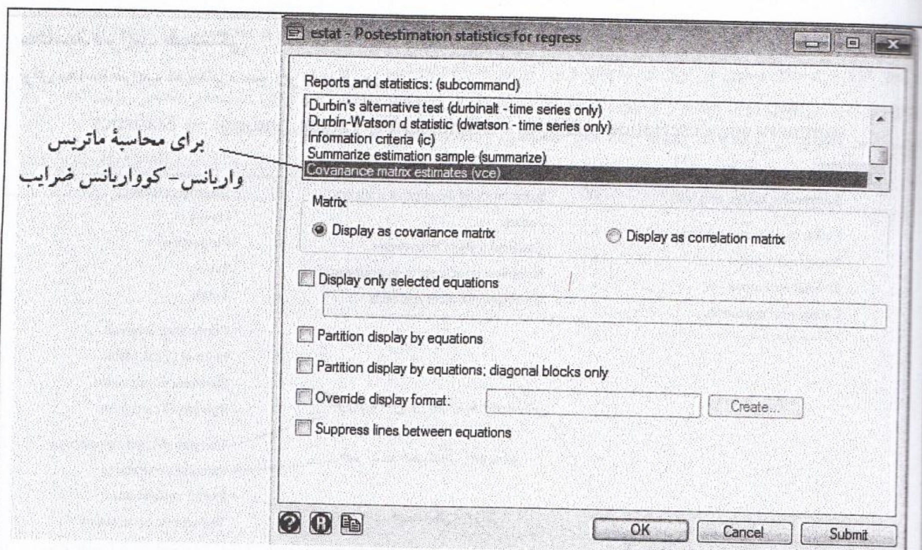


با انتخاب OK نتایج پیش‌بینی برون نمونه‌ای به‌دست می‌آید که می‌توان مقادیر آن را مشاهده و ترسیم نمود. برای رسم آن می‌توان از فرمان tsline ip ipff if t>1365 استفاده نمود.

محاسبهٔ ماتریس واریانس- کوواریانس ضرایب

بدین منظور مسیر زیر را دنبال می‌کنیم:

Report and Statistics → Postestimation → Statistics



نتایج عبارتند از:

```
Results

. estat vce

Covariance matrix of coefficients of regress model

        e(v) |        rr       _cons
-------------+------------------------
          rr |  105330.8
       _cons | 338563.27     6695678
.
```

کوواریانس ضرایب — واریانس ضریب rr — واریانس عرض از مبدأ

برآورد معادلات غیرخطی با Stata

در سطر فرمان معادلهٔ مورد نظر را وارد می‌کنیم. مثال‌های زیر نحوهٔ تخمین معادلات غیرخطی را نشان می‌دهد:

$Y=\alpha+\beta X \quad \Rightarrow$ nl (y={b0}+{b1}*x)

$Y=\alpha+\dfrac{\beta}{X} \quad \Rightarrow$ nl (y={b0}+{b1}/x)

$Y=\alpha+\beta X^{2} \quad \Rightarrow$ nl (y={b0}+{b1}*x^2)

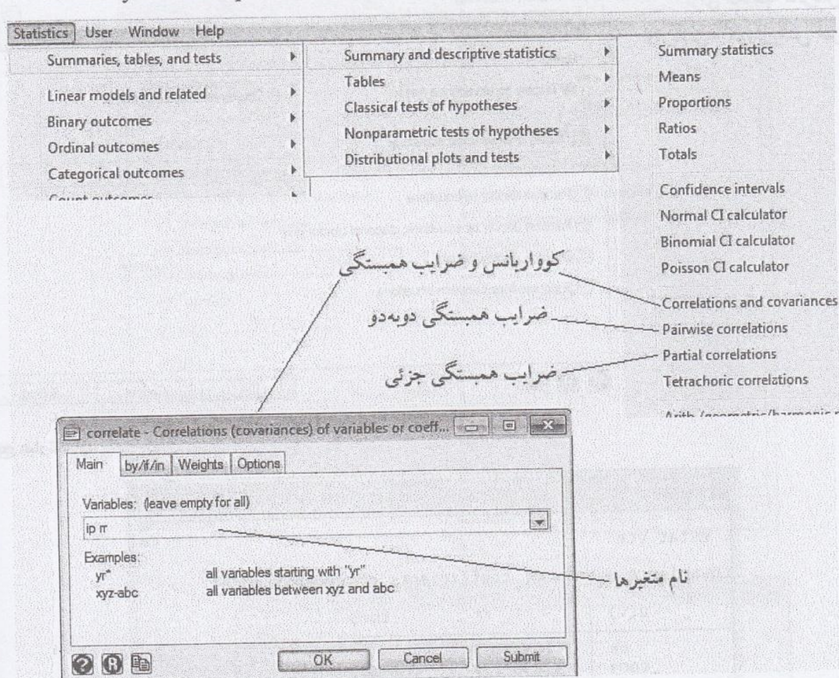
$Y=\alpha X^{\beta} \quad \Rightarrow$ nl (y={b0}*x^{b1})

$Y=\alpha+\beta X^{\theta} \quad \Rightarrow$ nl (y={b0}+{b1}*x^{b3})

محاسبهٔ ضرایب همبستگی

برای محاسبهٔ ضرایب همبستگی، مسیر زیر را دنبال می‌کنیم:

summary and description statistics → Summaries, Tables, and test → Statistics



نتایج عبارتند از:

```
Results
. correlate ip rr
(obs=21)

             |       ip       rr
-------------+------------------
          ip |   1.0000
          rr |  -0.5128   1.0000

.
```

لحاظ نکردن برخی از مشاهدات

اگر بخواهیم در تخمین معادلات و یا در سایر محاسبات، شرطی را قایل شویم می‌توان آن را با if لحاظ نمود. به عنوان مثال می‌خواهیم آن بخش از مشاهدات که y برابر صفر است، در نظر گرفته نشوند. بدین منظور از فرمان زیر استفاده می‌کنیم:

```
Drop if y==0
```

این فرمان موجب حذف آن بخش از مشاهدات می‌شود که $y=0$ است. همچنین می‌توان به جای صفر هر عدد دیگری را نیز درنظر گرفت. به جای مساوی می‌توان از کوچکتر و بزرگتر نیز استفاده نمود.

علاوه بر این، می‌توان روی معادله رگرسیون شرط گذاشت. مثلاً می‌خواهیم معادله را به ازای داده‌هایی تخمین بزنیم که $Z>0$ باشد. Z هر متغیری می‌تواند باشد، اعم از اینکه متغیر توضیحی، وابسته و یا حتی در معادله موجود نباشد.

```
reg y x  if z>=0
```

فصل سوم

رگرسیون چندمتغیره

۱-۳ مقدمه

در فصل دوم، رگرسیون ساده و مفاهیم اصلی آن تشریح گردید. در این فصل رگرسیون چندمتغیره را بررسی می‌کنیم که در حالت کلی شامل K متغیر توضیحی است. برای سادگی بحث، ابتدا حالت دو متغیره و سپس حالت کلی را بررسی خواهیم کرد.

۲-۳ رگرسیون دو متغیره: مفاهیم و فروض

در اینجا برای سادگی، بحث را با رگرسیون دو متغیره یعنی رگرسیونی که دارای دو متغیر توضیحی است، شروع می‌کنیم. بدین منظور معادله زیر را در نظر بگیرید:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{۳-۱}$$

که در آن $\alpha + \beta_1 X_{1i} + \beta_2 X_{2i}$ جزء غیرتصادفی یا معین و u_i جزء تصادفی یا اخلال می‌باشد.

در مورد u_i تمام فروض رگرسیون یک متغیره برقرار است و لذا برای Y_i داریم:

$$E\left(Y \middle| X_1, X_2\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{۳-۲}$$

$$\mathrm{var}\left(Y \middle| X_1, X_2\right) = \mathrm{var}(u_i \mid X_1, X_2) = \sigma^2 \tag{۳-۳}$$

یکی از مشکلات مدل‌های چندمتغیره، مسئله همخطی بین متغیرهای توضیحی X_1 و X_2 است. همخطی یا وابستگی خطی بین این دو متغیر را می‌توان به‌صورت زیر بیان نمود:

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} = 0 \tag{۳-۴}$$

اگر هیچ همخطی وجود نداشته باشد، آنگاه $\lambda_1 = \lambda_2 = 0$ است. ولی اگر همخطی، کامل باشد در این صورت λ_1 و λ_2 صفر نخواهند بود و لذا می‌توان (۳-۴) را به‌صورت زیر نوشت:

$$X_{2i} = -\frac{\lambda_2}{\lambda_1} X_{1i} \Rightarrow X_{2i} = aX_{1i} \quad ; \; a = -\frac{\lambda_2}{\lambda_1} \qquad (۳\text{-}۵)$$

حال اگر (۳-۵) را در (۳-۱) قرار دهیم، خواهیم داشت:

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{1i} + \beta_2 (aX_{1i}) + u_i \\ &= \alpha + (\beta_1 + \beta_2 a) X_{1i} + u_i \\ &= \alpha + \beta X_{1i} + u_i \quad ; \quad \beta = \beta_1 + \beta_2 a \end{aligned} \qquad (۳\text{-}۶)$$

بنابراین در صورت همخطی کامل، نمی‌توان ضرایب مدل (۳-۱) را برآورد نمود، بلکه فقط دو ضرایب α و β برآورد می‌شوند و امکان برآورد β_1 و β_2 وجود ندارد.

۳-۳ تخمین ضرایب رگرسیون دو متغیره و خواص آنها

برای تخمین ضرایب مدل رگرسیون دو متغیره، مانند رگرسیون یک متغیره، مقدار واقعی را با Y_i و مقدار تخمینی را با $\hat{Y}_i$ نشان می‌دهیم. لذا خطای تخمین عبارت است از:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i})$$

و مجموع مجذور خطاها عبارت است از:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2 \qquad (۳\text{-}۷)$$

ضرایب $\hat{\alpha}$، $\hat{\beta}_1$ و $\hat{\beta}_2$ باید به‌گونه‌ای تعیین شوند که مجموع مجذور خطا، حداقل شود:

$$\begin{aligned} \frac{\partial \sum e_i^2}{\partial \hat{\alpha}} &= -2\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \quad \text{یا} \quad \sum e_i = 0 \\ \frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} &= -2\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{1i} = 0 \quad \text{یا} \quad \sum e_i X_{1i} = 0 \\ \frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} &= -2\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{2i} = 0 \quad \text{یا} \quad \sum e_i X_{2i} = 0 \end{aligned} \qquad (۳\text{-}۸)$$

معادله اول را به‌صورت $\sum Y_i = n\hat{\alpha} + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i}$ نوشته و با تقسیم طرفین آن بر n رابطه $\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ را به‌دست می‌آوریم. اگر در معادلات دوم و سوم به‌جای $\hat{\alpha}$ قرار داده و آنها را بر حسب انحراف از میانگین بنویسیم، خواهیم داشت:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})X_{1i} = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})X_{2i} = 0$$

با ساده نمودن دو معادله فوق، خواهیم داشت:

$$\sum x_{1i} y_i - \hat{\beta}_1 \sum x_{1i}^2 - \hat{\beta}_2 \sum x_{1i} x_{2i} = 0 \quad (3-9)$$

$$\sum x_{2i} y_i - \hat{\beta}_1 \sum x_{1i} x_{2i} - \hat{\beta}_2 \sum x_{2i}^2 = 0 \quad (3-10)$$

که شکل ماتریسی آن عبارت است از:

$$\begin{bmatrix} \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{bmatrix} \quad (3-11)$$

با حل معادلات فوق، $\hat{\beta}_1$ و $\hat{\beta}_2$ به‌دست می‌آید.

مثال ۱-۳: داده‌های مربوط به سه متغیر X_1، X_2 و Y در جدول زیر ارائه شده است:

Y	X_1	X_2
۱۰	۲	۱۰۰
۱۲	۲	۱۰۰
۱۲	۴	۱۱۰
۱۴	۵	۱۲۰
۱۴	۵	۱۲۰
۱۵	۴	۱۴۰
۱۴	۶	۱۵۰
۱۶	۶	۱۵۰
۱۸	۸	۱۲۰
۲۰	۱۰	۱۵۰
۲۰	۱۰	۱۶۰
۲۲	۸	۱۶۰
۲۲	۱۰	۱۸۰
۲۴	۱۲	۲۰۰
۲۵	۱۲	۲۲۰
۲۵	۱۲	۲۲۰
۲۸	۱۴	۲۲۰
۲۹	۱۶	۲۴۰
۳۰	۱۶	۲۵۰
۳۰	۱۸	۲۵۰
۴۰۰	۱۸۰	۳۳۶۰

از جدول فوق $\overline{X}_1 = ۹$ ، $\overline{X}_2 = ۱۶۸$ و $\overline{Y} = ۲۰$ به‌دست می‌آید. نتایج حاصل از بقیه محاسبات به‌صورت زیر است:

$$\sum x_{1i}^2 = ۴۳۴ \quad , \quad \sum x_{2i}^2 = ۴۹۳۲۰$$
$$\sum y^2 = ۸۰۰ \quad , \quad \sum x_{1i}x_{2i} = ۴۴۰۰$$
$$\sum x_{1i}y_i = ۵۷۶ \quad , \quad \sum x_{2i}y_i = ۶۰۴۰$$

تخمین ضرایب عبارت است از:

$$\begin{bmatrix} ۴۳۴ & ۴۴۰۰ \\ ۴۴۰۰ & ۴۹۳۲۰ \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} ۵۷۶ \\ ۶۰۴۰ \end{bmatrix} \Rightarrow \hat{\beta}_1 = ۰/۸۹۶۰۵ \ , \ \hat{\beta}_2 = ۰/۰۴۲۵۳$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2 = ۲۰ - (۰/۸۹۶۰۵)(۹) - (۰/۰۴۲۵۳)(۱۶۸) = -۱۳/۲۰۹۵$$

بنابراین معادله رگرسیون عبارت است از:

$$\hat{Y}_i = -۱۳/۲۰۹۵ + ۰/۸۹۶۰۵X_{1i} + ۰/۰۴۲۵۳X_{2i}$$

از آنجا که $\hat{\beta}_1$ و $\hat{\beta}_2$ تابعی از Y و به تبع آن تابعی از u می‌باشند لذا دارای توزیع نرمال با امید ریاضی β_1 و β_2 هستند[1]. از طرف دیگر، واریانس $\hat{\beta}_1$ و $\hat{\beta}_2$، علاوه بر σ^2 به ضریب همبستگی

۱- برای اثبات این قضایا لازم است که $\hat{\beta}_1$ را به‌صورت زیر بنویسیم:

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i y_i \quad , \quad w_i = \frac{\sum x_{1i}x_{2i} - x_{2i}\sum x_{1i}x_{2i}}{\sum x_{1i}\sum x_{2i} - (\sum x_{1i}x_{2i})^2}$$

بدین ترتیب $\hat{\beta}_1$ ترکیب خطی از Y_i است که اگر به‌جای Y_i قرار دهیم، خواهیم داشت:

$$\hat{\beta}_1 = \sum w_i Y_i = \sum w_i(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i) = \beta_1 + \sum w_i u_i$$

بنابراین می‌توان به‌سادگی نتیجه گرفت که $E(\hat{\beta}_1) = \beta_1$ است. برای محاسبه واریانس $\hat{\beta}_1$ نیز به‌صورت زیر عمل می‌کنیم:

$$\text{var}(\hat{\beta}_1) = E(\hat{\beta}_1 - \beta_1) = E(\sum w_i u_i) = E(\sum w_i^2 u_i^2)$$

زیرا $E(u_i u_j) = ۰$ است.

$$\text{var}(\hat{\beta}_1) = \sum w_i^2 E(u_i^2) = \sum w_i^2 \sigma^2 = \sigma^2 \sum w_i^2$$

با محاسبهٔ $\sum w_i^2$ خواهیم داشت:

$$\text{var}(\hat{\beta}_1) = \sigma^2 \sum w_i^2 = \sigma^2 \frac{\sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} = \frac{\sigma^2}{(1 - r_{12}^2)\sum x_{1i}^2}$$

بین X_1 و X_2 نیز بستگی دارد. هر چه ضریب همبستگی X_1 و X_2 بیشتر باشد، واریانس‌ها نیز بزرگتر می‌شود و همچنان که در آزمون فرضیهٔ معنی‌دار بودن ضرایب رگرسیون خواهیم دید، ضرایب بی‌معنی خواهند شد. در صورت همبستگی کامل بین X_1 و X_2 ضریب همبستگی آنها برابر با ۱ است $(r_{12}^2 = 1)$ و در این حالت واریانس‌ها بی‌نهایت خواهند شد. اما اگر $r_{12}^2 = 0$ باشد دراین صورت واریانس $\hat{\beta}_1$ مشابه واریانس آن در معادله رگرسیون Y روی X_1 می‌باشد. همین بحث را برای X_2 نیز می‌توان مطرح نمود.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{(1-r_{12}^2)\sum x_{1i}^2}\right)$$

(۳-۱۲)

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{(1-r_{12}^2)\sum x_{2i}^2}\right)$$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left[\frac{1}{n}+\frac{(\bar{X}_1)^2\sum x_{2i}^2+(\bar{X}_2)^2\sum x_{1i}^2-2\bar{X}_1\bar{X}_2\sum x_{1i}x_{2i}}{(1-r_{12}^2)\sum x_{1i}^2 x_{2i}^2}\right]\right)$$

$$r_{12}^2 = X_2 \text{ و } X_1 \text{ مجذور ضریب همبستگی } = \frac{\left(\sum x_{1i}x_{2i}\right)^2}{\sum x_{1i}^2\sum x_{2i}^2}$$

۳-۴ ضریب تعیین و خطای معادلهٔ رگرسیون

برآورد واریانس u که با $\hat{\sigma}^2$ نشان داده می‌شود، واریانس معادلهٔ رگرسیون می‌باشد که مشابه رگرسیون یک متغیره می‌توان آن را محاسبه نمود. انحراف معیار($\hat{\sigma}$) نشان می‌دهد که به‌طور متوسط، مقادیر واقعی Y چقدر از مقادیر تخمینی انحراف دارند. طبق تعریف، واریانس معادله رگرسیون برابر با متوسط مجذور خطاها می‌باشد که از تقسیم RSS به درجه آزادی به‌دست می‌آید. نکته مهم در رگرسیون چندمتغیره این است که با افزایش تعداد متغیرهای توضیحی، معمولاً RSS کاهش و ESS افزایش می‌یابد. بدین ترتیب با افزایش تعداد متغیرهای توضیحی، قدرت توضیح‌دهندگی مدل و لذا مقدار R^2 افزایش می‌یابد.

اگر برای معادله $Y=\alpha+\beta X+u$ ضریب تعیین را با R_x^2 و برای معادله $Y=\alpha+\beta X+\gamma Z+u$ (که متغیر Z را اضافه کرده‌ایم) ضریب تعیین را با R_{xz}^2 نشان دهیم، رابطه بین R^2 ها عبارت است از[1]:

$$R_{xz}^2 = R_x^2 + (1-R_x^2)r_{yz.x}^2 \qquad (۳-۱۳)$$

که $r_{yz.x}^2$ ضریب همبستگی جزئی بین Y و Z است. بنابراین $R_{xz}^2 \geq R_x^2$ می‌باشد.

اما هر متغیر جدیدی که به مدل اضافه می‌شود از یک طرف R^2 را افزایش و از طرف دیگر درجه آزادی را کاهش می‌دهد. به همین منظور از معیار دیگری به‌نام « R^2 تعدیل‌شده» استفاده می‌شود که آن را با $\bar{R}^2$ نشان می‌دهند. در محاسبهٔ $\bar{R}^2$ اثر تعداد پارامترهای برآوردشده، تعدیل می‌شود به‌گونه‌ای که به‌جای تغییرات کل، از متوسط تغییرات استفاده می‌شود. بدیهی است که الزاماً $\bar{R}^2 \leq R^2$ است. همچنین ثابت می‌شود که برای هر متغیر جدیدی که به مدل اضافه می‌شود، اگر قدر مطلق t مربوطه، بزرگتر از ۱ باشد موجب افزایش $\bar{R}^2$ می‌شود. برای محاسبه σ^2، R^2 و $\bar{R}^2$ از روابط زیر می‌توان استفاده نمود که تا حدود زیادی مشابه با رگرسیون ساده می‌باشد:

$$\hat{\sigma}^2 = \frac{\sum(Y_i-\hat{Y}_i)^2}{n-K} = \frac{\sum e_i^2}{n-K} = \frac{RSS}{n-K} \qquad (۳-۱۴)$$

$$R^2 = \frac{ESS}{TSS} = 1-\frac{RSS}{TSS}$$

$$\bar{R}^2 = 1-\frac{RSS/(n-K)}{TSS/(n-1)} = 1-\frac{n-1}{n-K}\frac{RSS}{TSS} = 1-\frac{n-1}{n-K}(1-R^2) \qquad (۳-۱۵)$$

$$TSS = \sum y_i^2$$

$$ESS = \hat{\beta}_1\sum x_{1i}y_i + \hat{\beta}_2\sum x_{2i}y_i \qquad (۳-۱۶)$$

$$RSS = TSS - ESS$$

به خاطر تخمین K پارامتر، درجه آزادی برابر با $n-K$ است که در اینجا $K=3$ است.

مثال ۳-۲: در مثال ۳-۱ خطای معادله رگرسیون و ضریب تعیین را حساب کنید.

[1] Green, 2003, p. 34

$$TSS=\sum y_i^{۲}=۸۰۰$$
$$ESS=\hat{\beta}_{۱}\sum x_{۱i}y_i+\hat{\beta}_{۲}^{۲}\sum x_{۲i}y_i$$
$$=(۰/۸۹۶۰۵)(۵۷۶)+(۰/۰۴۲۵۳)(۶۰۴۰)=۷۷۳/۰۰۶$$
$$RSS=۸۰۰-۷۷۳/۰۰۶=۲۶/۹۹۴$$
$$\hat{\sigma}^{۲}=\frac{RSS}{n-۳}=\frac{۲۶/۹۹۴}{۲۰-۳}=۰/۹۹۹۷۷۸$$
$$R^{۲}=\frac{ESS}{TSS}=\frac{۷۷۳/۰۰۶}{۸۰۰}=۰/۹۶۶۳$$
$$\overline{R}^{۲}=۱-\frac{n-۱}{n-k}(۱-R^{۲})=۱-\frac{۲۰-۱}{۲۰-۳}(۱-۰/۹۶۶۳)=۰/۹۶۲۲$$

۵-۳ معیارهای اطلاعات[1]

در محاسبه $\overline{R}^{۲}$ دیدیم که این معیار در خصوص افزودن یک متغیر توضیحی مورد استفاده قرار گیرد، زیرا افزودن یک متغیر اضافی، از یک طرف RSS را کاهش و از طرف دیگر درجه آزادی را کاهش می‌دهد. در واقع نفع و ضرر حاصل از افزودن یک متغیر را لحاظ می‌کند. مشابه این، معیارهای دیگری معرفی شده‌اند که معروف به معیار اطلاعات هستند. این معیارها نیز نفع و ضرر حاصل از افزودن یک یا چند متغیر را لحاظ می‌کنند. بدین معنی که هم نفع حاصل از کاهش RSS و هم زیان حاصل از کاهش درجه آزادی (افزایش K) را در نظر می‌گیرند. این معیارها شامل معیار اطلاعات آکائیک (AIC)، معیار اطلاعات بیزین شوارتز (SIC) و معیار اطلاعات حنان- کوئین $(HQIC)$ می‌باشند:

$$AIC=\ln\hat{\sigma}^{۲}+\frac{۲K}{n}$$
$$SIC=\ln\hat{\sigma}^{۲}+\frac{K}{n}(\ln n) \quad (۳-۱۷)$$
$$HQIC=\ln\hat{\sigma}^{۲}+\frac{۲K}{n}\ln(\ln n)$$

که $\hat{\sigma}^{۲}=\frac{RSS}{n}$ می‌باشد. افزودن یک متغیر توضیحی به مدل موجب افزایش جمله اول (از طریق افزایش K) و موجب کاهش جمله دوم (از طریق کاهش $\hat{\sigma}^{۲}$) می‌شود. بنابراین، هر چقدر معیارهای اطلاعات کمتر شوند، معیار بهتری برای انتخاب مدل است. در هر سه معیار، جمله اول شامل ضریبی از $\frac{K}{n}$ است. این ضریب در معیار آکائیک کمتر از بقیه است، در حالی که در معیار

[1] Information criteria

حنان- کوئین بیشتر می‌باشد. در واقع معیار حنان- کوئین وزن بیشتری به $\frac{K}{n}$ می‌دهد و لذا جریمه بیشتری را به‌خاطر ورود یک متغیر در نظر می‌گیرد.

معیارهای اطلاعات در Eviews

هر معادله‌ای که با Eviews که تخمین می‌زنیم، خروجی آن هر سه معیار اطلاعات را ارائه می‌دهد. با مراجعه به جدول نتایج تخمین، مقدار معیارهای اطلاعات را می‌توان مشاهده نمود. اگر یک یا چند متغیر به مدل اضافه کنیم، آنگاه مقدار معیارهای اطلاعات تغییر خواهد کرد. اگر افزودن یک متغیر باعث کاهش معیار اطلاعات شود، در این صورت این متغیر را نگه می‌داریم و در غیر این صورت آن را کنار می‌گذاریم.

۶-۳ آزمون معنی‌دار بودن ضرایب معادله رگرسیون

مشابه رگرسیون یک متغیره، برای هر یک از ضرایب معادله رگرسیون می‌توان آزمون معنی‌دار بودن را انجام داد. همان طور که در رگرسیون یک متغیره دیدیم، اگر تعداد مشاهدات زیاد باشد، در صورتی که برای هر یک از ضرایب $|t|\geq ۲$ باشد، آن ضریب در سطح ۵ درصد معنی‌دار است. ولی برای نمونه‌هایی که حجم آنها تقریباً کمتر از ۲۰ باشد بایستی مقدار t را با عدد بحرانی (t جدول) مقایسه کنیم. مقدار t از تقسیم ضریب موردنظر (یعنی $\hat{\beta}_i$) به انحراف معیار آن (یعنی $SE(\hat{\beta}_i)=\sqrt{\text{var}(\hat{\beta}_i)}$) به‌دست می‌آید. معمولاً برای سادگی، مقادیر t را داخل پرانتز و در زیر هر یک از ضرایب معادله رگرسیون می‌نویسند.

مثال ۳-۳: در مثال ۲-۳ معنی‌دار بودن ضرایب رگرسیون را بررسی کنید:

ابتدا واریانس $\hat{\alpha}$، $\hat{\beta}_۱$ و $\hat{\beta}_۲$ را طبق روابط ۱۱-۳ حساب می‌کنیم:

$$r_{۱۲}^{۲}=\frac{(۴۴۰۰)^{۲}}{(۴۳۴)(۴۹۳۲۰)}=۰/۹۰۴۴۶۷$$

$$\text{var}(\hat{\beta}_۱)=\frac{۰/۹۹۹۷۷۸}{(۱-۰/۹۰۴۴۶۷)(۴۳۴)}=۰/۰۲۴۱۱۴ \quad , \quad SE(\hat{\beta}_۱)=۰/۱۵۵۳$$

$$\text{var}(\hat{\beta}_۲)=\frac{۰/۹۹۹۷۷۸}{(۱-۰/۹۰۴۴۶۷)(۴۹۳۲۰)}=۰/۰۰۰۲۱۲ \quad , \quad SE(\hat{\beta}_۲)=۰/۰۱۴۵۶۷$$

$$\text{var}(\hat{\alpha})=۰/۹۹۹۷۷۸\left(\frac{۱}{۲۰}+\frac{(۹)^{۲}(۴۹۳۲۰)+(۱۸۶)^{۲}(۴۳۴)-۲(۹)(۱۸۶)(۴۴۰۰)}{(۱-۰/۹۰۴۴۶۷)(۴۳۴)(۴۹۳۲۰)}\right)$$

$$=۲/۰۹۶۳۲۹$$

$$\hat{Y}_i = ۴/۷۹۱۲ + ۰/۸۹۶۰۵X_{۱i} + ۰/۰۴۲۵۳X_{۲i} \qquad R^۲ = ۰/۹۶۶۲$$
$$(۳/۳۱) \quad (۵/۷۷) \qquad (۲/۹۲) \qquad \overline{R}^۲ = ۰/۹۶۲۲$$

چون برای هر یک از ضرایب، قدر مطلق t از عدد بحرانی (یعنی۲/۰۸۶) بزرگتر است، همه ضرایب معنی‌دار هستند.

۷-۳ آزمون معنی‌دار بودن معادله رگرسیون (تحلیل واریانس)

تحلیل واریانس نشان می‌دهد که آیا کل معادله رگرسیون معنـی‌دار اسـت یـا خیـر. بـه عبـارت دیگر آیا تغییرات توضیح داده شده توسط متغیرهای توضیحی در مقایسه بـا تغییـرات توضـیح داده نشده (یعنی خطاها) معنی‌دار می‌باشد یا نه. مشابه رگرسیون یک متغیره، می‌توان این آزمون را با F انجام داد که بحث آن دقیقاً مشابه مبحث تحلیل واریانس است که در فصل دوم بررسی شد. بدین منظور بایستی فرضیه $H_۰ : \beta_۱ = \beta_۲ = ۰$ را آزمون کنیم.

$$F = \frac{S_E^۲}{S_R^۲} = \frac{S_E^۲}{\hat{\sigma}^۲} = \frac{ESS/(K-۱)}{RSS/(n-K)} = \frac{n-K}{K-۱}\frac{ESS}{RSS} = \frac{n-K}{K-۱}\frac{R^۲}{۱-R^۲} \qquad (۳-۱۸)$$

توجه شود که در اینجا K تعداد ضرایب تخمینی است که برابر با ۳ می‌باشد. هرگاه مقـدار F از مقدار بحرانیِ آن بزرگتـر باشـد (یعنـی $F \geq F_{\alpha,k-۱,n-K}$) بـدین معنـی اسـت کـه فرضـیه $H_۰$ رد می‌شود و رگرسیون معنادار است. بدیهی است که هرچه $R^۲$ بزرگتر باشد، F نیـز بزرگتـر خواهـد شد.

مثال ٤-۳: در مثال ۳-۳ معنی‌دار بودن معادله رگرسیون را انجام دهید.

بدین منظور فرضیه زیر را آزمون می‌کنیم:

$$H_۰ : \beta_۱ = \beta_۲ = ۰$$
$$H_۱ : \beta_i \neq ۰ \qquad i = ۱,۲$$

مقدار F برابر است با:

$$F = \frac{n-K}{K-۱}\frac{ESS}{RSS} = \frac{۲۰-۳}{۳-۱} \times \frac{۷۷۲/۹۸۱۷}{۲۷/۰۱۸۳} = ۲۴۳/۴۱$$

مقدار بحرانی F نیز برابر است با:

$$F_{\alpha,k-۱,n-K} = F_{۰/۰۵,۲,۱۷} = ۳/۵ \Rightarrow (F \geq ۳/۵)$$

چون مقدار F در ناحیه بحرانی قرار دارد، لذا فرضیه H_0 رد می‌شود و رگرسیون معنی‌دار است.

۳-۸ تحلیل نتایج رگرسیون دو متغیره

فرض کنید معادله رگرسیون به‌صورت زیر باشد:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \qquad (۳-۱۹)$$

β_1 تأثیر X_1 بر Y است و نشان می‌دهد که اگر X_1 یک واحد تغییر کند، Y به اندازه β_1 تغییر خواهد کرد. حال فرض کنید به‌جای (۱۹-۳) به اشتباه معادله زیر را برآورد کنیم:

$$Y_i = b_0 + b_1 X_{1i} + \varepsilon_{1i} \qquad (۳-۲۰)$$

b_1 تأثیر X_1 بر Y است و ε_{1i} جملهٔ اخلال را نشان می‌دهد. اما سؤال این است که چه تفاوتی بین b_1 و β_1 است.

توجه داریم که $\hat{\beta}_1$ تخمین‌زنندهٔ نااریب (بدون تورش) از β_1 می‌باشد. در حالی که ما به‌جای آن $\hat{b}_1$ را برآورد نموده‌ایم که $\hat{b}_1$ برآوردکنندهٔ b_1 است نه β_1. می‌توان نشان داد که $\hat{b}_1$ و $\hat{\beta}_1$ با هم برابر نیستند. فقط در یک حالت این دو برابرند و آن زمانی است که X_1 و X_2 هیچ ارتباطی با هم نداشته باشند و ضریب همبستگی آنها (r_{12}) صفر باشد. اما عملاً چنین نیست و ضریب همبستگی X_1 و X_2 معمولاً مخالف صفر است. بدین منظور معادله زیر را در نظر بگیرید:

$$X_{2i} = B_0 + B_1 X_{1i} + \varepsilon_{2i} \qquad (۳-۲۱)$$

ضریب B_1 بیانگر اثر X_1 بر X_2 است. اگر X_1 و X_2 کاملاً مستقل باشند، آنگاه B_1 برابر صفر خواهد بود و در این صورت b_1 و β_1 نیز برابرند.

حال(۲۱-۳) را در(۱۹-۳) جایگزین می‌کنیم:

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{1i} + \beta_2 (B_0 + B_1 X_{1i} + \varepsilon_{2i}) + u_i \\ &= (\alpha + \beta_2 B_0) + (\beta_1 + \beta_2 B_1) X_{1i} + (u_i + \beta_2 \varepsilon_{2i}) \end{aligned} \qquad (۳-۲۲)$$

مقایسه (۲۲-۳) با (۲۰-۳) نشان می‌دهد که $b_0 = \alpha + \beta_2 B_0$، $b_1 = \beta_1 + \beta_2 B_1$ و $\varepsilon_{1i} = u_i + \beta_2 \varepsilon_{2i}$ است. بنابراین وقتی معادله (۲۰-۳) را به‌جای معادله (۱۹-۳) برآورد کنیم به این معنی است که به‌جای β_1، ضریب b_1 را برآورد نموده‌ایم که قطعاً با β_1 برابر نیست:

$$b_{۱} = \beta_{۱} + \beta_{۲} B_{۱} \tag{۳-۲۳}$$

حال تفسیر $b_{۱}$ چگونه است. $b_{۱}$ شامل دو اثر است: یکی اثر $X_{۱}$ بر Y با ثابت نگه داشتن $X_{۲}$ که معادل با $\beta_{۱}$ می‌باشد. دیگری تأثیر غیرمستقیم $X_{۱}$ بر Y است که از طریق تـأثیر $X_{۱}$ بـر $X_{۲}$ روی Y اعمال می‌شود و برابر با $\beta_{۲}B_{۱}$ می‌باشد. لذا $b_{۱}$ را می‌توان اثر ناخالص $X_{۱}$ بر Y دانست:

(۳-۲۴) **اثرغیرمستقیم $X_{۱}$ از طریق $X_{۲}$ بر Y + اثر خالص یا مستقیم $X_{۱}$ بر Y = اثرناخالص $X_{۱}$ بر Y**

$b_{۱} = \beta_{۱} + \beta_{۲}B_{۱}$

بنابراین اگر مدل واقعی به‌صورت (۱۹-۳) باشد و ما به اشتباه $X_{۲}$ را از معادلـه حـذف کنـیم و معادله (۲۰-۳) برآورد نمائیم. در این صورت اثر $X_{۱}$ بر Y را به‌صورت غیر واقعی برآورد نموده‌ایم که مقدار خطا برابر با $\beta_{۲}B_{۱}$ است. به عبارت دیگر، $b_{۱}$ اثر $X_{۱}$ بـر Y بـدون ثابـت نگـه‌داشـتن $X_{۲}$ است و $\beta_{۱}$ اثر $X_{۱}$ بر Y با ثابت نگه‌داشتن $X_{۲}$ می‌باشد.

مثال ۵-۳: در مثال ۱-۳ اثر ناخالص، مستقیم و غیرمستقیم $X_{۱}$ بر Y را به‌دست آوردید.

در اینجا باید $b_{۱} = \beta_{۱} + \beta_{۲}B_{۱}$ را برآورد نماییم. اثـر ناخـالص $X_{۱}$ بـر Y از معادلـه زیـر به‌دست می‌آید:

$$\hat{Y}_i = \hat{b}_{.} + \hat{b}_{۱} X_{۱i}$$

$$\hat{b}_{۱} = \frac{\sum x_{۱i} y_i}{\sum x_{۱i}^{۲}} = \frac{۵۷۶}{۴۳۴} = ۱/۳۲۷۲$$

$$\hat{b}_{.} = \overline{Y} - \hat{b}_{۱}\overline{X}_{۱} = ۲۰ - (۱/۳۲۷۲)(۹) = ۸/۰۵۵۳$$

$$\hat{Y}_i = ۱/۳۲۷۲ + ۸/۰۵۵۳ X_{۱i}$$

اثر خالص $X_{۱}$ بر Y در مثال ۱-۳ برابر با $\hat{\beta}_{۱} = ۰/۸۹۶۰۵$ بـه‌دسـت آمـد. از طـرف دیگـر برای تعیین $B_{۱}$ داریم:

$$\hat{X}_{۲i} = \hat{B}_{.} + \hat{B}_{۱} \hat{X}_{۱i}$$

$$\hat{B}_{۱} = \frac{\sum x_{۱i} x_{۲i}}{\sum x_{۱i}^{۲}} = \frac{۴۴۰۰}{۴۳۴} = ۱۰/۱۳۸۲$$

$$\hat{B}_{.} = \overline{X}_{۲} - \hat{B}_{۱}\overline{X}_{۱} = ۱۶۸ - (۱۰/۱۳۸۲)(۹) = ۷۶/۷۵۵۸$$

$$\hat{X}_{۲i} = ۷۶/۷۵۵۸ + ۱۰/۱۳۸۲ X_{۱i}$$

بنـابراین اثـر ناخـالص (کـل) $X_۱$ بـر Y برابـر بـا $\hat{b}_۱ = ۱/۳۲۷۲$ و اثـر خـالص برابـر بـا $\hat{\beta}_۱ = ۰/۸۹۶۰۵$ و اثـر غیـر مـستقیم برابـر بـا $\hat{\beta}_۲\hat{B}_۱ = (۰/۰۴۲۵۳)(۱۰/۱۳۸۲) = ۰/۴۳۱۲$ می‌باشد.

نتایج فوق را می‌توان با استفاده از معادله (۹-۳) نیز به‌دست آورد. بدین منظور $\hat{\beta}_۱$ را به‌صورت زیر می‌نویسیم:

$$\sum x_{۱i}y_i - \hat{\beta}_۱\sum x_{۱i}^۲ - \hat{\beta}_۲\sum x_{۱i}x_{۲i} = ۰ \Rightarrow \hat{\beta}_۱ = \frac{\sum x_{۱i}y_i}{\sum x_{۱i}^۲} - \hat{\beta}_۲\frac{\sum x_{۱i}x_{۲i}}{\sum x_{۱i}^۲}$$

ملاحظه می‌شود که در فرمول فوق، $\hat{b}_۱ = \frac{\sum x_{۱i}y_i}{\sum x_{۱i}^۲}$ و $\hat{B}_۱ = \frac{\sum x_{۱i}x_{۲i}}{\sum x_{۱i}^۲}$ و لذا $\hat{\beta}_۱$ برابر است با:

$$\hat{\beta}_۱ = \hat{b}_۱ - \hat{\beta}_۲\hat{B}_۱$$

طبق رابطه فوق، اثر خالص $X_۱$ بر Y، یعنی $(\hat{\beta}_۱)$ برابر با اثر ناخالص $(\hat{b}_۱)$ منهای اثر $X_۱$ بر Y از طریق $X_۲$ می‌باشد.

۹-۳ تحلیل همبستگی چندمتغیره

ضرایب همبستگی بیانگر روابط خطـی بـین متغیرهـا هـستند. بـه‌عنـوان مثـال در رگرسـیون دو متغیره، $r_{۱۲}$ ضریب همبستگی بین $X_۱$ و $X_۲$ را نـشان مـی‌دهـد. ضـریب همبـستگی بـین $X_۱$ و Y را با $r_{۱y}$ و بین $X_۲$ و Y را با $r_{۲y}$ نشان می‌دهیم. هـر یـک از ایـن ضـرایب همبـستگی را بـا اسـتفاده از فرمول «ضریب همبستگی ساده» می‌توان محاسبه نمود.

حال با توجه به اینکه ۳ متغیر $X_۱$، $X_۲$ و Y را داریم، تا چه اندازه $r_{۱y}$ بیـانگر همبـستگی واقعـی بین $X_۱$ و Y است. در حالی که $X_۲$ نیز با هر دو متغیرها همبستگی دارد. این بحث دقیقاً مـشابه آن چیزی است که در رابطه با رگرسیون و اثر واقعی $X_۱$ بر Y بررسی نمودیم.

$r_{۱y}$ بیانگر همبستگی بین $X_۱$ و Y در حالتی است که هیچ توجهی به $X_۲$ نداشته باشیم. در واقع به‌طور ضمنی فرض کرده‌ایم که $X_۲$ ثابت است، در حـالی کـه چنـین نیـست. بنـابراین $r_{۱y}$ تـصویر غیرواقعی از همبستگی بین $X_۱$ و Y را ارائه می‌دهد. بدین ترتیب لازم است که اثـر $X_۲$ را بـه‌طـور کامل حذف کنیم تا بدین ترتیب بتوان همبستگی واقعی بین $X_۱$ و Y را به‌دست آورد. این ضریب

همبستگی را «ضریب همبستگی جزئی» بین X_1 و Y می‌گوییم که با $r_{1y.2}$ نشان می‌دهیم که به معنی «ضریب همبستگی بین X_1 و Y با ثابت نگه داشتن X_2» است.

$r_{1y.2}$: ضریب همبستگی جزئی بین X_1 و Y با ثابت نگه داشتن X_2
$r_{2y.1}$: ضریب همبستگی جزئی بین X_2 و Y با ثابت نگه داشتن X_1
$r_{12.y}$: ضریب همبستگی جزئی بین X_1 و X_2 با ثابت نگه داشتن Y

برای محاسبه «ضرایب همبستگی جزئی» روابط نسبتاً ساده‌ای وجود دارد که آنها را می‌توان بر حسب «ضرایب همبستگی ساده» بیان نمود.

$$r_{1y.2} = \frac{r_{1y} - r_{2y} r_{12}}{\sqrt{\left(1 - r_{2y}^2\right)\left(1 - r_{12}^2\right)}}$$

$$r_{2y.1} = \frac{r_{2y} - r_{1y} r_{12}}{\sqrt{\left(1 - r_{1y}^2\right)\left(1 - r_{12}^2\right)}} \qquad (۳-۲۵)$$

$$r_{12.y} = \frac{r_{12} - r_{1y} r_{2y}}{\sqrt{\left(1 - r_{1y}^2\right)\left(1 - r_{2y}^2\right)}}$$

تفسیر ضرایب همبستگی جزئی را می‌توان با استفاده از روابط (۲۵-۳) ارائه نمود. به‌عنوان مثال در ضریب همبستگی جزئی $r_{1y.2}$ حتی اگر ضریب همبستگی ساده صفر باشد ($r_{1y} = 0$)، ضریب همبستگی جزئی $r_{1y.2}$ صفر نخواهد بود، زیرا:

$$r_{1y} = 0 \Rightarrow r_{1y.2} = -\frac{r_{2y} r_{12}}{\sqrt{(1 - r_{2y}^2)(1 - r_{12}^2)}}$$

فقط در صورتی ضریب همبستگی جزئی صفر می‌شود که r_{12} یا r_{2y} نیز صفر شود، یعنی X_1 با X_2 و همچنین Y با X_2 رابطه‌ای نداشته باشد.

حال اگر $r_{12} = 0$ باشد در این صورت همبستگی جزئی بین X_1 و Y به‌صورت زیر است:

$$r_{12} = 0 \quad \Rightarrow \quad r_{1y.2} = \frac{r_{1y}}{\sqrt{1 - r_{2y}^2}}$$

اگر علاوه بر $r_{12} = 0$، ضریب همبستگی ساده بین X_2 و Y نیز صفر باشد ($r_{2y} = 0$)، در این صورت ضریب همبستگی جزئی و ضریب همبستگی ساده برابر خواهند بود:

$$r_{12} = r_{2y} = 0 \Rightarrow r_{1y.2} = r_{1y}$$

اما بین ضریب تعیین و ضرایب همبستگی نیز روابطی وجود دارد که عبارتند از:

$$R^2 = \frac{r_{1y}^2 + r_{2y}^2 - 2r_{1y}r_{2y}r_{12}}{1 - r_{12}^2} \tag{۳-۲۶}$$
$$= r_{1y}^2 + (1 - r_{1y}^2)r_{2y.1}^2 = r_{2y}^2 + (1 - r_{2y}^2)r_{1y.2}^2$$

مثال ۶-۳: در مثال ۱-۳ ضرایب همبستگی ساده و جزئی را حساب کنید:

ضرایب همبستگی ساده عبارتند از:

$$r_{12} = \frac{4400}{\sqrt{(434)(49320)}} = 0/951 \qquad r_{12}^2 = 0/904$$

$$r_{1y} = \frac{576}{\sqrt{(434)(800)}} = 0/977 \qquad r_{1y}^2 = 0/954$$

$$r_{2y} = \frac{6040}{\sqrt{(49320)(800)}} = 0/962 \qquad r_{2y}^2 = 0/925$$

و ضرایب همبستگی جزئی، طبق روابط (۶۳-۲۰) عبارتند از:

$$r_{1y.2} = \frac{0/977 - (0/962)(0/951)}{\sqrt{(1-0/925)(1-0/904)}} = 0/736$$

$$r_{2y.1} = \frac{0/962 - (0/977)(0/951)}{\sqrt{(1-0/954)(1-0/904)}} = 0/498$$

$$r_{12.y} = \frac{0/951 - (0/977)(0/962)}{\sqrt{(1-0/954)(1-0/925)}} = 0/191$$

۱۰-۳ مدل K متغیره

در حالت کلی معادله زیر را در نظر بگیرید که دارای K متغیر توضیحی می‌باشد:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i \tag{۳-۲۷}$$

در این معادله β_1 عرض از مبدأ می‌باشد که در واقع ضریب متغیر X_1 است و X_1 نیز برداری با عناصر واحد می‌باشد. حال مدل فوق را به‌صورت ماتریسی می‌نویسیم:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{K1} \\ 1 & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2n} & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \qquad (۳-۲۸)$$

بنابراین مدل (۲۸-۳) به‌طور خلاصه عبارت است از:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad (۳-۲۹)$$

$\mathbf{y}$ و $\mathbf{u}$ بردارهای ستونی با ابعاد $n\times 1$ هستند. $\mathbf{X}$ ماتریس $n\times K$ و $\boldsymbol{\beta}$ بردار ستونی $K\times 1$ می‌باشد.

در اینجا نیز بایستی $\boldsymbol{\beta}$ به‌گونه‌ای برآورد شود که که مجموع مجذور خطا (RSS) حداقل شود:

$$RSS = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} \quad , \quad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

$\mathbf{e}$ بردار ستونی خطاها و $\mathbf{e}'$ ترانسپوز آن می‌باشد. با جایگذاری به‌جای $\mathbf{e}$، خواهیم داشت:

$$\begin{aligned} RSS &= \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned} \qquad (۳-۳۰)$$

برای حداقل شدن $\mathbf{e}'\mathbf{e}$، نسبت به $\hat{\boldsymbol{\beta}}$ مشتق می‌گیریم:

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{X} = \mathbf{o} \qquad (۳-۳۱)$$

با حل این معادله، $\hat{\boldsymbol{\beta}}$ به‌دست می‌آید:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) \qquad (۳-۳۲)$$

$\hat{\boldsymbol{\beta}}$ تخمین‌زنندهٔ OLS برای $\boldsymbol{\beta}$ است. این تخمین‌زننده چون تابع خطی از $\mathbf{u}$ است، لذا دارای امید ریاضی $\boldsymbol{\beta}$ و واریانس $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ می‌باشد. بدین منظور ابتدا به‌جای Y قرار می‌دهیم:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \end{aligned}$$

امید ریاضی $\hat{\boldsymbol{\beta}}$ عبارت است از:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}) = \boldsymbol{\beta} \quad , \quad E(\mathbf{u}) = \mathbf{o} \end{aligned}$$

واریانس $\hat{\boldsymbol{\beta}}$ عبارت است از:

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] \\ &= E\left\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\right\} \\ &= E\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

$E(\mathbf{u}\mathbf{u}')$ ماتریس واریانس-کوواریانس $\mathbf{u}$ است که آن را با $\mathbf{\Sigma}$ نشان می‌دهیم:

$$E(\mathbf{u}\mathbf{u}') = \mathbf{\Sigma} = E\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}\begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix}$$

$$= E\begin{bmatrix} u_1^2 & u_1u_2 & \dots & u_1u_n \\ u_1u_2 & u_2^2 & \dots & u_2u_n \\ \vdots & \vdots & & \vdots \\ u_1u_n & u_2u_n & \dots & u_n^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}_n$$

زیرا $E(u_i^2) = \sigma^2$ و $E(u_iu_j) = 0$ است. بنابراین، واریانس $\hat{\boldsymbol{\beta}}$ برابر است با:

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned} \qquad (۳-۳۳)$$

بدین ترتیب واریانس $\hat{\beta}_k$ برابر با حاصل‌ضرب σ^2 در عنصر kام قطر اصلی ماتریس $(\mathbf{X}'\mathbf{X})^{-1}$ می‌باشد. سایر عناصر نیز کواریانس بین ضرایب را نشان می‌دهد. به‌طور کلی، ماتریس $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ را ماتریس واریانس- کوواریانس ضرایب می گویند.

ماهیت این ماتریس در مثال ۷-۳ نشان داده شده است. از آنجا که σ^2 مجهول است، لذا از تخمین آن استفاده می کنیم که برابر است با:

$$\hat{\sigma}^2 = \frac{RSS}{n-K} = \frac{\mathbf{e}'\mathbf{e}}{n-K} \qquad (۳-۳۴)$$

$n-K$ برابر با درجه آزادی است، زیرا K پارامتر را برآورد کرده‌ایم.

مثال ۷-۳: فرض کنید که بر اساس داده‌های ۱۵ سال برای سه متغیر Y، $X_۲$ و $X_۳$ مقادیر زیر به‌دست آمده است:

$$(\mathbf{XX}')^{-۱}=\begin{bmatrix} ۲/۰ & ۳/۵ & -۱/۰ \\ ۳/۵ & ۱/۰ & ۶/۵ \\ -۱/۰ & ۶/۵ & ۴/۳ \end{bmatrix},\quad \mathbf{X'y}=\begin{bmatrix} -۳/۰ \\ ۲/۲ \\ ۰/۶ \end{bmatrix}$$

$$\mathbf{e'e}=۱۰/۹۶$$

بر اساس این مشاهدات، $\hat{\boldsymbol{\beta}}$ برابر است با:

$$\hat{\boldsymbol{\beta}}=(\mathbf{X'X})^{-۱}(\mathbf{X'y})=\begin{bmatrix} ۱/۱ \\ -۴/۴ \\ ۱۹/۸۸ \end{bmatrix}$$

تخمین واریانس u برابر است با:

$$\hat{\sigma}^۲=\frac{\mathbf{e'e}}{n-K}=\frac{۱۰/۹۶}{۱۵-۳}=۰/۹۱$$

بنابراین، واریانس $\hat{\boldsymbol{\beta}}$ عبارت است از:

$$\text{var}(\hat{\boldsymbol{\beta}})=\hat{\sigma}^۲(\mathbf{X'X})^{-۱}=۰/۹۱\begin{bmatrix} ۲/۰ & ۳/۵ & -۱/۰ \\ ۳/۵ & ۱/۰ & ۶/۵ \\ -۱/۰ & ۶/۵ & ۴/۳ \end{bmatrix}$$

$$=\begin{bmatrix} ۱/۸۲ & ۳/۱۹ & -۰/۹۱ \\ ۳/۱۹ & ۰/۹۱ & ۵/۹۲ \\ -۰/۹۱ & ۵/۹۲ & ۳/۹۱ \end{bmatrix}=\begin{bmatrix} \text{var}(\hat{\beta}_۱) & \text{cov}(\hat{\beta}_۱,\hat{\beta}_۲) & \text{cov}(\hat{\beta}_۱,\hat{\beta}_۳) \\ \text{cov}(\hat{\beta}_۲,\hat{\beta}_۱) & \text{var}(\hat{\beta}_۲) & \text{cov}(\hat{\beta}_۲,\hat{\beta}_۳) \\ \text{cov}(\hat{\beta}_۳,\hat{\beta}_۱) & \text{cov}(\hat{\beta}_۳,\hat{\beta}_۲) & \text{var}(\hat{\beta}_۳) \end{bmatrix}$$

بدین ترتیب واریانس ضرایب عبارت است از:

$$\text{var}(\hat{\beta}_۱)=۱/۸۲\ ,\quad \text{var}(\hat{\beta}_۲)=۰/۹۱\ ,\quad \text{var}(\hat{\beta}_۳)=۳/۹۱$$

معادله رگرسیون عبارت است از:

$$\hat{Y}_i=۱/۱۰-۴/۴۰X_{۲i}+۱۹/۸۸X_{۳i}$$
$$(۰/۸۱)\quad(-۴/۶)\quad(۱۰/۰۵)$$

برآورد رگرسیون چند متغیره با Eviews — فایل data3

برای برآورد معادله رگرسیون، مشابه رگرسیون یک متغیره می‌توان عمل نمود. در پنجره فرمان، معادله مورد نظر را به‌صورت LS Y C X1 X2 وارد می‌کنیم. در اینجا Y متغیر وابسته، C عرض از مبدأ و X1 و X2 متغیرهای توضیحی هستند. بعد از وارد نمودن عبارت مذکور، با زدن کلید Enter نتایج تخمین در پنجره Equation ارائه می‌شود. تمامی نتایج مشابه رگرسیون یک متغیره می‌باشد.

فرض کنید که می‌خواهیم رگرسیون سرمایه‌گذاری (I) را روی نرخ بهره حقیقی (RR)، تغییرات درآمـد ملـی بـدون نفـت (dY) و سرمایه‌گذار در سال قبل(۱-)I برآورد کنیم، بدین منظور از دستور زیر استفاده می‌کنیم:

LS I C RR dY I(-1)

با زدن کلید Enter نتایج به‌صورت زیر نشان داده می‌شود:

Equation: UNTITLED Workfile: DATA5::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: I
Method: Least Squares
Date: 09/23/11 Time: 19:17
Sample (adjusted): 1352 1386
Included observations: 35 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5742.205	5095.942	1.126819	0.2685
RR	-246.7890	200.1556	-1.232986	0.2268
DYNO	1.187917	0.158345	7.502104	0.0000
I(-1)	0.819466	0.052327	15.66052	0.0000

R-squared	0.947025	Mean dependent var	90623.63
Adjusted R-squared	0.941899	S.D. dependent var	35779.97
S.E. of regression	8624.471	Akaike info criterion	21.06981
Sum squared resid	2.31E+09	Schwarz criterion	21.24756
Log likelihood	-364.7216	Hannan-Quinn criter.	21.13117
F-statistic	184.7287	Durbin-Watson stat	1.655497
Prob(F-statistic)	0.000000		

جدول فوق نشان می‌دهد که عرض از مبدأ، معنادار نیست زیرا برای آن $t = ۱/۱۲$ است. ضریب نرخ بهره حقیقی برابر با ۲۴۶/۸- است که علامت منفی دارد و مطابق با تئوری است، اما از لحاظ آماری معنادار نیست. ضریب تغییرات درآمد ملی، هم مثبت و هم معنادار است. متغیر سرمایه‌گذاری تأخیری نیز دارای ضریب مثبت و معنادار است. $R^۲$ برابر با ۰/۹۴۷ و $\overline{R}^۲$ نیـز برابـر بـا ۰/۹۴۲ می‌باشد. مقدار F نیز برابر با ۱۸۴/۷ است که کاملاً معنادار می‌باشد.

ماتریس $\mathbf{X'X}$ نقـش مهمـی در بـرآورد $\boldsymbol{\beta}$ دارد. بـرای بررسـی خـواص مـاتریس $\mathbf{X'X}$، ابتـدا متغیرها را بر حسب انحراف از میانگین می‌نویسیم. بدین منظور معادله (۳-۲۷) را جمـع زده و بـر n تقسیم می‌کنیم:

$$\sum Y_i = n\beta_۱ + \beta_۲ \sum X_{۲i} + \ldots + \beta_K \sum X_{Ki} + \sum u_i$$

$$\overline{Y} = \beta_۱ + \beta_۲ \overline{X}_۲ + \ldots + \beta_K \overline{X}_K + \overline{u} \qquad (۳-۳۵)$$

حال معادله فوق را از (۲۷-۳) کم می کنیم:

$$Y_i - \overline{Y} = \beta_۲(X_{۲i} - \overline{X}_۲) + ... + \beta_K(X_{Ki} - \overline{X}_K) + (u_i - \overline{u})$$

متغیرهایی را که بر حسب انحراف از میانگین هستند با حروف کوچک نشان می‌دهیم:

$$y_i = \beta_۲ x_{۲i} + ... + \beta_K x_{Ki} + v_i \quad (۳-۲۷')$$

شکل ماتریسی معادله فوق عبارت است از (مشابه معادله ۲۹-۳):

$$\mathbf{y}_* = \mathbf{X}_* \boldsymbol{\beta}_* + \mathbf{v} \quad (۳-۲۹')$$

$\boldsymbol{\beta}_*$ یک بردار $(K-۱)\times ۱$ و $\mathbf{X}_*$ نیز $n\times(K-۱)$ می‌باشد که شامل داده‌های $x_۲$ تا x_K است. در اینجا نیز مجموع مجذور خطاها $(\mathbf{e'e})$ را حساب کرده و با مشتق‌گیری نسبت به $\hat{\boldsymbol{\beta}}_*$، تخمین پارامترها را به‌دست می‌آوریم:

$$\mathbf{e} = \mathbf{y}_* - \hat{\mathbf{y}}_* = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}_*$$

$$\mathbf{e'e} = (\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}_*)'(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}_*) = \mathbf{y}'_*\mathbf{y}_* - ۲\mathbf{y}'_*\mathbf{X}_*\hat{\boldsymbol{\beta}}_* + \hat{\boldsymbol{\beta}}'_*(\mathbf{X}'_*\mathbf{X}_*)\hat{\boldsymbol{\beta}}_*$$

$$\frac{\partial \mathbf{e'e}}{\hat{\boldsymbol{\beta}}_*} = \mathbf{o} \Rightarrow \hat{\boldsymbol{\beta}}_* = (\mathbf{X}'_*\mathbf{X}_*)^{-۱}\mathbf{X}'_*\mathbf{y}_* \quad (۳-۳۳')$$

با تخمین $\hat{\boldsymbol{\beta}}_*$، تخمین $\hat{\beta}_۲$ تا $\hat{\beta}_K$ به‌دست می‌آید. $\hat{\beta}_۱$ نیز از رابطه زیر تعیین می‌شود:

$$\hat{\beta}_۱ = \overline{Y} - \hat{\beta}_۲\overline{X}_۲ - ... - \hat{\beta}_K\overline{X}_K$$

همان‌طور که اشاره شد، ماتریس $\mathbf{X}_*$ در اینجا شامل داده‌های $x_۲$ تا x_K می‌باشد:

$$\mathbf{X}_* = \begin{bmatrix} x_{۲۱} & x_{۳۱} & \cdots & x_{K۱} \\ x_{۲۲} & x_{۳۲} & \cdots & x_{K۲} \\ \vdots & \vdots & & \vdots \\ x_{۲n} & x_{۳n} & \cdots & x_{Kn} \end{bmatrix} \quad (۳-۳۶)$$

ماتریس $\mathbf{X}'_*\mathbf{X}_*$ عبارت است از:

$$\mathbf{X}'_*\mathbf{X}_* = \begin{bmatrix} \sum x_{2i}^2 & \sum x_{2i}x_{3i} & \cdots & \sum x_{2i}x_{Ki} \\ \sum x_{3i}x_{2i} & \sum x_{3i}^2 & \cdots & \sum x_{3i}x_{Ki} \\ \vdots & \vdots & & \vdots \\ \sum x_{Ki}x_{2i} & \sum x_{Ki}x_{3i} & \cdots & \sum x_{Ki}^2 \end{bmatrix} \qquad (۳-۳۷)$$

اگر هر یک از عناصر ماتریس $\mathbf{X}'_*\mathbf{X}_*$ را بر n تقسیم کنیم آن‌گاه ماتریسی به‌دست می‌آید که قطر اصلی آن واریانس Xها و سایر عناصر، کوواریانس بین Xها را نشان می‌دهد:

$$\frac{1}{n}\mathbf{X}'_*\mathbf{X}_* = \begin{bmatrix} S_2^2 & S_{23} & \cdots & S_{2K} \\ S_{32} & S_3^2 & \cdots & S_{3K} \\ \vdots & \vdots & & \vdots \\ S_{K2} & S_{K3} & \cdots & S_K^2 \end{bmatrix} \qquad (۳-۳۸)$$

S_i^2 واریانس X_i و S_{ij} کوواریانس X_i و X_j می‌باشد.

حال برای سادگی، بحث را بر اساس رگرسیون دو متغیره ادامه می‌دهیم. در این صورت $\mathbf{X}'_*\mathbf{X}_*$ و $\mathbf{X}'_*\mathbf{y}_*$ عبارتند از:

$$\mathbf{X}'_*\mathbf{X}_* = \begin{bmatrix} \sum x_{2i}^2 & \sum x_{2i}x_{3i} \\ \sum x_{3i}x_{2i} & \sum x_{3i}^2 \end{bmatrix} \; ; \; \mathbf{X}'_*\mathbf{y}_* = \begin{bmatrix} \sum x_{2i}y_i \\ \sum x_{3i}y_i \end{bmatrix}$$

$(\mathbf{X}'_*\mathbf{X}_*)^{-1}$ نیز عبارت است از:

$$(\mathbf{X}'_*\mathbf{X}_*)^{-1} = \frac{1}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i}x_{3i})^2} \begin{bmatrix} \sum x_{3i}^2 & -\sum x_{2i}x_{3i} \\ -\sum x_{3i}x_{2i} & \sum x_{2i}^2 \end{bmatrix}$$

$$= \frac{1}{(1-r_{23}^2)\sum x_{2i}^2 \sum x_{3i}^2} \begin{bmatrix} \sum x_{3i}^2 & -\sum x_{2i}x_{3i} \\ -\sum x_{3i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \qquad (۳-۳۹)$$

$$r_{23}^2 = \frac{(\sum x_{2i}x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2}$$

r_{23} ضریب همبستگی X_2 و X_3 است.

براساس ماتریس $(\mathbf{X}'_*\mathbf{X}_*)^{-1}$ می‌توان سه حالت را بررسی نمود:

الف) اگر بین X_2 و X_3 رابطه خطی کامل باشد، در این صورت $r_{23}^2 = 1$ خواهد بود و لذا $(\mathbf{X}'_*\mathbf{X}_*)^{-1}$ قابل محاسبه نیست و امکان برآورد ضرایب وجود ندارد. در این حالت به‌عنوان

مثال $X_{3i}=aX_{2i}$ است، یعنی یکی از ستون‌های ماتریس $\mathbf{X}_*$ ضریبی از ستون دیگر است. در این صورت ما فقط راجع به یکی از متغیرهای توضیحی دارای اطلاعات هستیم و در مورد دیگری هیچ اطلاعاتی نداریم و لذا فقط یکی از ضرایب آنها (ضریب X_2 یا ضریب X_3) قابل برآورد است. به‌عبارت دیگر اطلاعات ما برای برآورد دوضریب، کافی نیست. اگر $X_{3i}=aX_{2i}$ باشد، خواهیم داشت:

$$\begin{aligned} Y_i &= \beta_1+\beta_2X_{2i}+\beta_3(aX_{2i})+u_i \\ &= \beta_1+(\beta_2+\beta_3a)X_{2i}+u_i \\ &= \beta_1+\alpha_2X_{2i}+u_i \end{aligned}$$

بنابراین فقط ضریب X_2 قابل برآورد است و امکان برآورد ضریب X_3 وجود ندارد.

ب) اگر X_2 و X_3 مستقل خطی باشند، در این صورت $r_{23}^2=0$ است که بدیهی است $\sum x_{2i}x_{3i}=0$ می‌باشد. در این حالت، $(\mathbf{X}_*'\mathbf{X}_*)^{-1}$ و برآورد ضرایب X_2 و X_3 عبارت است از:

$$\hat{\boldsymbol{\beta}}_*=\begin{bmatrix}\hat\beta_2\\ \hat\beta_3\end{bmatrix}=\frac{1}{\sum x_{2i}^2\sum x_{3i}^2}\begin{bmatrix}\sum x_{3i}^2 & 0\\ 0 & \sum x_{2i}^2\end{bmatrix}\begin{bmatrix}\sum x_{2i}y_i\\ \sum x_{3i}y_i\end{bmatrix}=\begin{bmatrix}\dfrac{\sum x_{2i}y_i}{\sum x_{2i}^2}\\ \dfrac{\sum x_{3i}y_i}{\sum x_{3i}^2}\end{bmatrix} \tag{3-40}$$

نتایج فوق نشان می‌دهد که تخمین ضرایب β_2 و β_3 مشابه تخمین آنها در معادلات رگرسیون یک متغیره است، یعنی β_2 از معادله $Y_i=\alpha+\beta_2X_{2i}+u_i$ و β_3 از معادله $Y_i=\alpha+\beta_3X_{3i}+u_i$ به‌دست می‌آید.

ج) اگر X_2 و X_3 رابطه خطی غیر کامل داشته باشند به‌گونه‌ای که $0<r_{23}^2<1$ باشد، در این حالت تخمین $\hat\beta_2$ و $\hat\beta_3$ به هم وابسته هستند. بدین منظور معادلات زیر را داریم:

$$(\mathbf{X}_*'\mathbf{X}_*)\hat{\boldsymbol{\beta}}_*=\mathbf{X}_*'\mathbf{y}_*\Rightarrow\begin{bmatrix}\sum x_{2i}^2 & \sum x_{2i}x_{3i}\\ \sum x_{2i}x_{3i} & \sum x_{3i}^2\end{bmatrix}\begin{bmatrix}\hat\beta_2\\ \hat\beta_3\end{bmatrix}=\begin{bmatrix}\sum x_{2i}y_i\\ \sum x_{3i}y_i\end{bmatrix} \tag{3-41}$$

$$\hat\beta_2\sum x_{2i}^2+\hat\beta_3\sum x_{2i}x_{3i}=\sum x_{2i}y_i\ \Rightarrow\ \hat\beta_2=\frac{\sum x_{2i}y_i}{\sum x_{2i}^2}-\frac{\sum x_{2i}x_{3i}}{\sum x_{2i}^2}\hat\beta_3 \tag{3-42}$$

$$\hat\beta_2\sum x_{2i}x_{3i}+\hat\beta_3\sum x_{3i}^2=\sum x_{3i}y_i\ \Rightarrow\ \hat\beta_3=\frac{\sum x_{3i}y_i}{\sum x_{3i}^2}-\frac{\sum x_{2i}x_{3i}}{\sum x_{3i}^2}\hat\beta_2 \tag{3-43}$$

نتایج فوق نشان می‌دهد که تخمین هر ضریب، وابسته به دیگری است. این بدان معنا است که اثر $X_۲$ بر Y معادل با $\beta_۲$ است که از معادله $Y_i = \beta_۱ + \beta_۲ X_{۲i} + \beta_۳ X_{۳i} + u_i$ تعیین می‌شود که در آن $X_۳$ وجود دارد. حال اگر $\beta_۲$ را از معادله $Y_i = \beta_۱ + \beta_۲ X_{۲i} + u_i$ به‌دست آوریم، دارای تورش خواهد بود و با $\beta_۲$ واقعی برابر نخواهد بود.

اگر از (۴۳-۳) به‌جای $\hat{\beta}_۳$ در (۴۲-۳) قرار دهیم، خواهیم داشت:

$$\hat{\beta}_۲ = \frac{\sum x_{۳i}^۲ \sum x_{۲i} y_i - \sum x_{۲i} x_{۳i} \sum x_{۳i} y_i}{\sum x_{۲i}^۲ \sum x_{۳i}^۲ - (\sum x_{۲i} x_{۳i})^۲} \qquad (۳-۴۴)$$

رابطه (۴۴-۳) را می‌توان با توجه به تعریف $r_{۲۳}^۲$ به‌صورت زیر نوشت:

$$\hat{\beta}_۲ = \frac{۱}{۱ - r_{۲۳}^۲} \left(\underbrace{\frac{\sum x_{۲i} y_i}{\sum x_{۲i}^۲}}_{\hat{b}_۲} - \underbrace{\frac{\sum x_{۲i} x_{۳i}}{\sum x_{۲i}^۲}}_{\hat{\alpha}_۲} \underbrace{\frac{\sum x_{۳i} y_i}{\sum x_{۳i}^۲}}_{\hat{b}_۳} \right) \qquad (۳-۴۵)$$

$\hat{b}_۲$، $\hat{\alpha}_۲$ و $\hat{b}_۳$ به ترتیب از معادلات (۴۶-۳)، (۴۷-۳)و (۴۸-۳) به‌دست می‌آیند:

$$Y_i = b_۱ + b_۲ X_{۲i} + u_i \qquad (۳-۴۶)$$

$$X_{۳i} = \alpha_۱ + \alpha_۲ X_{۲i} + \varepsilon_i \qquad (۳-۴۷)$$

$$Y_i = a_۱ + b_۳ X_{۳i} + v_i \qquad (۳-۴۸)$$

رابطه (۴۵-۳) نشان می‌دهد که تخمین ضریب $X_۲$ وابسته به رابطه $X_۲$ و $X_۳$ است که این وابستگی توسط ضریب همبستگی آنها (یعنی $r_{۲۳}^۲$) و همچنین توسط $\hat{\alpha}_۲$ توصیف می‌شود.

۱۱-۳ رگرسیون مقید و تخمین‌زننده‌های مقید

به‌عنوان یک مثال ساده فرض کنید که در معادله رگرسیون $Y_i = \beta_۱ + \beta_۲ X_{۲i} + \beta_۳ X_{۳i} + u_i$ با قید $\beta_۲ + \beta_۳ = ۱$ مواجه‌ایم. این یک رگرسیون مقید است که ضرایب آن از مسئله زیر به‌دست می‌آید:

$$\begin{aligned} \min \ & \sum e_i^۲ = \sum (Y_i - \hat{\beta}_۱ - \hat{\beta}_۲ X_{۲i} - \hat{\beta}_۳ X_{۳i})^۲ \\ st. \ & \hat{\beta}_۲ + \hat{\beta}_۳ = ۱ \end{aligned} \qquad (۳-۴۹)$$

برای حل این مسئله، تابع لاگرانژ را تشکیل می‌دهیم:

$$L=\sum(Y_i-\hat{\beta}_1-\hat{\beta}_2X_{2i}-\hat{\beta}_3X_{3i})^2+\lambda(1-\hat{\beta}_2-\hat{\beta}_3) \quad (۳-۵۰)$$

حال نسبت به ضرایب مشتق گرفته و برابر صفر قرار می‌دهیم:

$$\begin{aligned}
\partial L/\partial\beta_1 &= -2\sum(Y_i-\hat{\beta}_1-\hat{\beta}_2X_{2i}-\hat{\beta}_3X_{3i})=0\\
\partial L/\partial\beta_2 &= -2\sum(Y_i-\hat{\beta}_1-\hat{\beta}_2X_{2i}-\hat{\beta}_3X_{3i})X_{2i}-\lambda=0\\
\partial L/\partial\beta_3 &= -2\sum(Y_i-\hat{\beta}_1-\hat{\beta}_2X_{2i}-\hat{\beta}_3X_{3i})X_{3i}-\lambda=0\\
\partial L/\partial\lambda &= 1-\hat{\beta}_2-\hat{\beta}_3=0
\end{aligned} \quad (۳-۵۱)$$

از معادله اول، $\hat{\beta}_1=\bar{Y}-\hat{\beta}_2\bar{X}_2-\hat{\beta}_3\bar{X}$ به‌دست می‌آید. به‌جای $\hat{\beta}_1$ در معادله دوم و سوم قرار داده و نتایج را به‌صورت زیر ساده می‌کنیم:

$$\begin{aligned}
&\sum(y_i-\hat{\beta}_2x_{2i}-\hat{\beta}_3x_{3i})X_{2i}-\lambda=0\\
&\sum(y_i-\hat{\beta}_2x_{2i}-\hat{\beta}_3x_{3i})X_{3i}-\lambda=0
\end{aligned} \quad (۳-۵۲)$$

y_i، x_{2i} و x_{3i} برحسب انحراف از میانگین هستند. حال اگر به‌جای $\hat{\beta}_3$ از $\hat{\beta}_3=1-\hat{\beta}_2$ در معادلات (۳-۵۲) قرار دهیم، خواهیم داشت:

$$\begin{aligned}
&\sum(y_i-\hat{\beta}_2(x_{2i}-x_{3i})-x_{3i})X_{2i}-\lambda=0\\
&\sum(y_i-\hat{\beta}_2(x_{2i}-x_{3i})-x_{3i})X_{3i}-\lambda=0
\end{aligned} \quad (۳-۵۳)$$

با حل معادلات فوق، $\hat{\beta}_2$ به‌دست می‌آید[1]:

$$\hat{\beta}_2=\frac{\sum(x_{2i}-x_{3i})(y_i-x_{3i})}{\sum(x_{2i}-x_{3i})^2} \quad (۳-۵۴)$$

و $\hat{\beta}_3$ نیز برابر است با:

$$\hat{\beta}_3=1-\hat{\beta}_2=\frac{\sum(x_{3i}-x_{2i})(y_i-x_{2i})}{\sum(x_{2i}-x_{3i})^2} \quad (۳-۵۵)$$

البته می‌توان به‌جای روش فوق، ابتدا از $\beta_3=1-\beta_2$ در معادله رگرسیون قرار داده و سپس ضرایب را برآورد نمود:

---

[1] توجه شود که رابطه $\sum x_{2i}X_{2i}=\sum x_{2i}^2$ و $\sum x_{2i}X_{3i}=\sum x_{2i}x_{3i}$ برقرار است. مشابه این روابط برای سایر متغیرها نیز برقرار است.

$$Y_i = \beta_1 + \beta_2 X_{2i} + (1-\beta_2)X_{3i} + u_i$$
$$= \beta_1 + \beta_2(X_{2i} - X_{3i}) + X_{3i} + u_i$$

$$Y_i - X_{3i} = \beta_1 + \beta_2(X_{2i} - X_{3i}) + u_i \qquad (۳-۵۶)$$

با تخمین معادله (۵۶-۳)، تخمین‌زنندهٔ β_1 و β_2 به‌دست می‌آید که مشابه قبل است. سپس β_3 نیز از $\beta_3 = 1 - \beta_2$ به‌دست می‌آید.

حال بحث فوق را در حالت کلی به‌کار می‌بریم. بدین منظور ابتدا رگرسیون و قیدها را به‌صورت ماتریسی می‌نویسیم:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$$
$$\mathbf{R\beta} = \mathbf{r} \qquad (۳-۵۷)$$

بردارهای $\mathbf{y}$، $\mathbf{u}$ و $\boldsymbol{\beta}$ و ماتریس X مشابه (۲۹-۳) می‌باشند. معادله دوم، قیدها را نشان می‌دهد. به‌عنوان مثال می‌تواند بیانگر قید $\beta_2 + \beta_3 = 1$ باشد که به‌صورت $[1 \quad 1]\begin{bmatrix}\beta_2 \\ \beta_3\end{bmatrix} = 1$ نوشته می‌شود. به‌هر حال، R یک ماتریس $m \times K$ و r یک بردار $m \times 1$ است (m تعداد قیدها است). تخمین مقید از مسئله زیر به‌دست می‌آید:

$$\min \sum e_i^2 = \mathbf{e'e} = (\mathbf{y} - \mathbf{X\hat{\beta}})'(\mathbf{y} - \mathbf{X\hat{\beta}}) = \mathbf{y'y} + \mathbf{\beta'X'X\beta} - 2\mathbf{\beta'X'y}$$
$$st. \ \mathbf{R\beta} = \mathbf{r} \qquad (۳-۵۸)$$

با تشکیل تابع لاگرانژ و مشتق‌گیری نسبت به $\hat{\boldsymbol{\beta}}$ و بردار ضرایب لاگرانژ (در اینجا $\boldsymbol{\lambda}$ بردار ستونی $m \times 1$ است) تخمین مقید ضرایب به‌دست می‌آید (ضرایب لاگرانژ را برابر با $2\boldsymbol{\lambda}$ در نظر می‌گیریم):

$$L = \mathbf{y'y} + \mathbf{\hat{\beta}'X'X\hat{\beta}} - 2\mathbf{\hat{\beta}'X'y} + 2\boldsymbol{\lambda}'(\mathbf{r} - \mathbf{R\hat{\beta}}) \qquad (۳-۵۹)$$

$$\partial L / \partial \hat{\boldsymbol{\beta}} = 2\mathbf{X'X\hat{\beta}} - 2\mathbf{X'y} - 2\mathbf{R'\lambda} = \mathbf{o} \qquad (۳-۶۰)$$

$$\partial L / \partial \boldsymbol{\lambda} = -2(\mathbf{R\hat{\beta}} - \mathbf{r}) = \mathbf{o} \qquad (۳-۶۱)$$

با حل معادلات (۶۰-۳) و (۶۱-۳) تخمین مقید به‌دست می‌آید که آن را با $\hat{\boldsymbol{\beta}}_R$ نشان می‌دهیم:

$$\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_{UR} + (\mathbf{X'X})^{-1}\mathbf{R'}[\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R'}]^{-1}(\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\beta}}_{UR}) \qquad (۳-۶۲)$$

$\hat{\boldsymbol{\beta}}_{UR} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ تخمین غیرمقید است.

امید ریاضی و واریانس $\hat{\boldsymbol{\beta}}_R$ عبارت است از:

$$E(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{\beta}$$
$$\text{var}(\hat{\boldsymbol{\beta}}_R) = \sigma^2(\mathbf{X'X})^{-1} - \sigma^2(\mathbf{X'X})^{-1}\mathbf{R'}[\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R'}]^{-1}\mathbf{R}(\mathbf{X'X})^{-1} \quad (۳-۶۳)$$

۱۲-۳ آزمون محدودیت‌ها

در برخی موارد علاوه بر آزمون معنی‌دار بودن ضرایب، لازم است آزمون‌های دیگری را در مورد ضرایب انجام دهیم. آزمون معنی‌دار بودن ضرایب که با t صورت می‌گیرد به معنی آزمون فرضیه $\beta_i = 0$ است. همچنین F که در تحلیل واریانس مورد استفاده قرار می‌گیرد برای آزمون فرضیه $\beta_2 = \beta_3 = ... = \beta_K = 0$ می‌باشد که به معنی آزمون معنادار بودن کل معادله رگرسیون است. علاوه بر این می‌توان فرضیه‌های دیگری نیز برای ضرایب مطرح و آزمون نمود. به‌عنوان مثال فرضیه $\beta_i = a$ که a یک مقدار ثابت است. یا آزمون فرضیه $\beta_2 + \beta_3 = 1$ و یا هر فرضیه دیگری راجع به ضرایب را می‌توان آزمون نمود. برخی از این فرضیه‌ها یا محدودیت‌ها نسبتاً ساده هستند که می‌توان برای آزمون آنها از t استفاده نمود. به‌عنوان مثال می‌توان فرضیه $\beta_2 = 5$ را با استفاده از t آزمون نمود:

$$t = \frac{\hat{\beta}_2 - 5}{\sqrt{\text{var}(\hat{\beta}_2)}} \quad (۳-۶۴)$$

همچنین می‌توان فرضیه $\beta_2 = \beta_3$ را به‌صورت $\beta_2 - \beta_3 = 0$ نوشت و سپس آن را با آماره t آزمون نمود:

$$t = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\sqrt{\text{var}(\hat{\beta}_2 - \hat{\beta}_3)}} \quad (۳-۶۵)$$

$\text{var}(\hat{\beta}_2 - \hat{\beta}_3)$ برابر با $\text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_3) - 2\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$ است.

علاوه بر این، فرضیه $\beta_2 + \beta_3 = 1$ را می‌توان با استفاده از t به‌صورت زیر انجام داد:

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - 1}{\sqrt{\text{var}(\hat{\beta}_2 + \hat{\beta}_3)}} \quad (۳-۶۶)$$

$\text{var}(\hat{\beta}_2 + \hat{\beta}_3)$ برابر با $\text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_3) + 2\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$ است.

روش عمومی برای آزمون محدودیت‌ها را می‌توان بر اساس آماره F ارائه نمود. فرض کنید رگرسیون K متغیره زیر را داریم:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_K X_{Kt} + u_t \qquad (3-67)$$

از بخش ۵-۳ به‌خاطر داریم که برای آزمون $\beta_2 = \beta_3 = \ldots = \beta_k = 0$ از F تحت عنوان تحلیل واریانس استفاده نمودیم. حال می‌توان با استفاده از F هر فرضیه دیگری را که در رابطه با ضرایب باشد، آزمون نمود. بدین منظور مدل (۶۷-۳) را رگرسیون غیرمقید[1] می‌نامیم، زیرا هیچ محدودیتی روی ضرایب آن اعمال نشده است. برای این مدل، ضریب تعیین را با R^2_{UR} و مجموع مجذور باقیمانده‌ها را با RSS_{UR} نشان می‌دهیم. حال می‌توان هر گونه محدودیتی را روی ضرایب اعمال کرده و سپس با در نظر گرفتن این محدودیتها، مدل مذکور را مجدداً برآورد می‌کنیم که به آن رگرسیون مقید می‌گوییم. ضریب تعیین این مدل را با R^2_R و مجموع مجذور باقیمانده‌ها را با RSS_R نشان می‌دهیم. توجه داریم که $RSS_{UR} \le RSS_R$ است، زیرا خطاهای رگرسیون مقید،[2] بیشتر از رگرسیون غیرمقید می‌باشد. اگر تفاضل $RSS_R - RSS_{UR}$ بزرگ باشد بدان معنا است که محدودیتهای مورد نظر معنی‌دار هستند و در غیر این صورت، معنادار نیستند. در چنین شرایطی می‌توان از F استفاده نمود که بیانگر معنادار بودن یا معنادار نبودن محدودیتهای مورد نظر می‌باشد. بدیهی است که اگر محدودیتهای اعمال شده واقعاً معنادار باشند، رگرسیونهای مقید و غیرمقید تفاوت قابل توجهی خواهند داشت و لذا تفاضل RSSها بزرگ می‌شود که موجب بزرگ شدن F شده و فرضیه H_0 (وجود محدودیت‌ها) رد می‌شود. در واقع بدان معنا است که این محدودیت‌ها، قدرت توضیح‌دهندگی مدل را به‌طرز محسوسی کاهش می‌دهند.

$$F = \frac{\frac{RSS_R - RSS_{UR}}{m}}{\frac{RSS_{UR}}{n-K}} = \frac{n-K}{m}\frac{RSS_R - RSS_{UR}}{RSS_{UR}} \qquad (3-68)$$

که m تعداد محدودیت‌ها است. به‌عنوان مثال، $\beta_2 + \beta_3 = 1$ بیانگر وجود یک محدودیت است. اما فرضیه‌های $\beta_2 + \beta_3 = 1$ و $\beta_5 = \beta_6$، بیانگر وجود دو محدویت است. همچنین آزمون فرضیه

[1] Unrestricted
[2] Restricted

$\beta_2=\beta_4=\beta_5=0$ بیانگر سه محدودیت است، زیرا سه محدودیت به‌صورت $\beta_2=0$، $\beta_4=0$ و $\beta_5=0$ داریم.

از آنجا که رابطه $R^2=1-\frac{RSS}{TSS}$ برقرار است، لذا می‌توان F را بر حسب R^2 نوشت:

$$F=\frac{n-K}{m}\frac{R_{UR}^2-R_R^2}{1-R_{UR}^2} \tag{۳-۶۹}$$

به‌جای F می‌توان از χ^2 نیز استفاده نمود. از آنجا که $RSS=\sum e_t^2$ را می‌توان به‌صورت $\frac{RSS}{\sigma^2}=\sum(\frac{e_t-\bar{e}}{\sigma})^2$ نوشت ($\bar{e}=0$)، لذا عبارت $\frac{e_t-\bar{e}}{\sigma}$ توزیع نرمال استاندارد و مجذور آن توزیع χ^2 با درجه آزادی ۱ دارد. اما مجموع آن دارای توزیع χ^2 با درجه آزادی $n-K$ است. همچنین $\frac{RSS_R}{\sigma^2}$ و $\frac{RSS_{UR}}{\sigma^2}$ نیز توزیع χ^2 دارند که درجه آزادی اولی برابر با $n-(K-m)$ و دومی برابر با $n-K$ می‌باشد، زیرا در اولی فقط $K-m$ پارامتر و در دومی K پارامتر را برآورد می‌کنیم. تفاضل $\frac{RSS_R-RSS_{UR}}{\sigma^2}$ نیز توزیع χ^2 با درجه آزادی $[n-(K-m)]-(n-K)=m$ دارد. اگر مقدار χ_m^2 بزرگ باشد، مشابه با بزرگ بودن F است.[1]

در حالت کلی، برای آزمون محدودیت $R\beta=r$ نیز از نسبت F استفاده می‌کنیم[2]. بدین منظور ابتدا خطاهای غیرمقید و مقید را حساب می‌کنیم:

$$\begin{aligned}\mathbf{e}_{UR}&=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{UR}\\ \mathbf{e}_R&=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_R\end{aligned} \tag{۳-۷۰}$$

خطاهای مقید را به‌صورت زیر می‌نویسیم:

$$\begin{aligned}\mathbf{e}_R&=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_R=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_R+\mathbf{X}\hat{\boldsymbol{\beta}}_{UR}-\mathbf{X}\hat{\boldsymbol{\beta}}_{UR}\\ &=(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{UR})-\mathbf{X}(\hat{\boldsymbol{\beta}}_R-\hat{\boldsymbol{\beta}}_{UR})\\ &=\mathbf{e}_{UR}-\mathbf{X}(\hat{\boldsymbol{\beta}}_R-\hat{\boldsymbol{\beta}}_{UR})\end{aligned} \tag{۳-۷۱}$$

حال مجموع مجذور خطاها (RSS) را حساب می‌کنیم:

---

[1] این روش، معروف به آزمون والد است که جزئیات آن در فصل هفتم ارائه شده است.

[2] سایر آزمون‌ها در فصل هفتم معرفی شده‌اند.

$$RSS_R = \mathbf{e}'_R\mathbf{e}_R = \mathbf{e}'_{UR}\mathbf{e}_{UR} + (\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR}) \qquad (۳-۷۲)$$

با استفاده از معادله فوق، تفاوت بین RSS های مقید و نامقید را به صورت زیر می‌نویسیم:

$$RSS_R - RSS_{UR} = \mathbf{e}'_R\mathbf{e}_R - \mathbf{e}'_{UR}\mathbf{e}_{UR} = (\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR}) \qquad (۳-۷۴)$$

از $\hat{\boldsymbol\beta}_R = \hat{\boldsymbol\beta}_{UR} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\hat{\boldsymbol\beta}_{UR})$ به‌جای $\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR}$ در (۳-۷۴) قرار می‌دهیم:

$$\begin{aligned} RSS_R - RSS_{UR} &= \mathbf{e}'_R\mathbf{e}_R - \mathbf{e}'_{UR}\mathbf{e}_{UR} = (\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR}) \\ &= (\mathbf{r} - \mathbf{R}\hat{\boldsymbol\beta}_{UR})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\hat{\boldsymbol\beta}_{UR}) \end{aligned} \qquad (۳-۷۵)$$

حال برای آزمون فرضیه $H_0: \mathbf{R}\boldsymbol\beta = \mathbf{r}$ نسبت F را مشابه (۳-۶۸) تشکیل می‌دهیم:

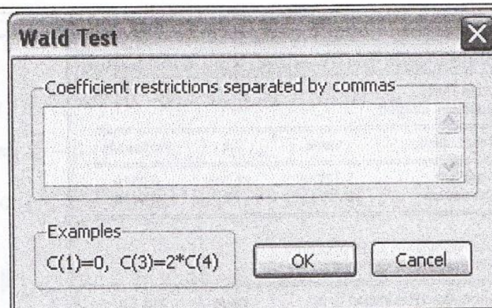
$$\begin{aligned} F_{m,n-K} &= \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-K)} \\ &= \frac{n-K}{m}\frac{(\mathbf{e}'_R\mathbf{e}_R - \mathbf{e}'_{UR}\mathbf{e}_{UR})}{\mathbf{e}'_{UR}\mathbf{e}_{UR}} \\ &= \frac{n-K}{m}\frac{(\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol\beta}_R - \hat{\boldsymbol\beta}_{UR})}{\mathbf{e}'_{UR}\mathbf{e}_{UR}} \\ &= \frac{n-K}{m}\frac{(\mathbf{r} - \mathbf{R}\hat{\boldsymbol\beta}_{UR})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\hat{\boldsymbol\beta}_{UR})}{\mathbf{e}'_{UR}\mathbf{e}_{UR}} \end{aligned} \qquad (۳-۷۶)$$

بنابراین، برای محاسبهٔ F فقط لازم است که رگرسیون غیرمقید را برآورد کرده و F را حساب می‌کنیم. اگر واقعاً محدودیتها بی‌تأثیر باشند آنگاه به‌ازای تخمین‌های غیرمقید نیز برقرار خواهند بود و لذا $\mathbf{r} - \mathbf{R}\hat{\boldsymbol\beta}$ به صفر نزدیک می‌شود که به‌دنبال آن مقدار F نیز کوچک شده و در ناحیه بحرانی قرار نمی‌گیرد و فرضیه H_0 رد نمی‌شود. این بدان معنا است که محدودیت‌های $\mathbf{R}\boldsymbol\beta = \mathbf{r}$ برقرار هستند.

آزمون محدودیت‌ها با Eviews: آزمون والد

فایل data3

در پنجره Equation (نتایج تخمین)، برای آزمون محدودیت‌ها، از منوی View گزینه Coefficient diagnostics و سپس گزینه Wald Test-Coefficient Restrictions را انتخاب می‌کنیم. در این صورت پنجره‌ای به نام Wald Test باز می‌شود.



در این پنجره، محدودیتهای مورد نظر را وارد می‌کنیم. به‌عنوان مثال اگر محدودیتهای $\beta_2 = 0$ ، $\beta_4 + \beta_5 = 1$ و $\beta_3 = -\beta_6$ را داشته باشیم، آنها را به‌صورت زیر وارد می‌کنیم:

$$C(2)=0, C(4)+C(5)=1, C(3)=-C(6)$$

اگر در رگرسیون سرمایه‌گذاری، ضریب نرخ بهره را برابر صفر قرار دهیم، سپس در پنجره wald test عبارت C(2)=0 را نوشته و OK را انتخاب می کنیم. نتایج آزمون در پنجره زیر نشان داده می‌شود:

Equation: UNTITLED Workfile: DATA5::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
t-statistic	-1.232986	31	0.2268
F-statistic	1.520254	(1, 31)	0.2268
Chi-square	1.520254	1	0.2176

Null Hypothesis: C(2)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(2)	-246.7890	200.1556

Restrictions are linear in coefficients.

در پنجره فوق، t-statistic بیانگر مقدار آماره t برای ضریب نرخ بهره در معادله اصلی (سرمایه‌گذاری) است (یعنی C(2)) که وضعیت معنادار بودن یا بودن آن نشان می‌دهد. چون مقدار F و χ^2 کوچک هستند (زیرا احتمال مربوطه بزرگتر از ۰/۰۵ است)، لذا محدودیت مورد نظر معنادار نیست و C(2) برابر با صفر است.

حال اگر هم زمان محدودیت C(2)=0 و C(1)=0 را اعمال نماییم (یعنی عرض از مبدأ و ضریب نرخ بهره را برابر صفر قراردهیم) نتایج آزمون والد به‌صورت زیر می‌باشد:

Equation: UNTITLED Workfile: DATA5::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	3.127557	(2, 31)	0.0579
Chi-square	6.255115	2	0.0438

Null Hypothesis: C(1)=0, C(2)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(1)	5742.205	5095.942
C(2)	-246.7890	200.1556

Restrictions are linear in coefficients.

در اینجا مقدار F و $\chi^{۲}$ بزرگ هستند (زیرا سطح احتمال مربوطه نزدیک به ۰/۰۵ و یا کوچکتر هستند) و لذا محدودیت‌های مورد نظر، معنادار هستند. مقدار آماره والد برابر با ۶/۲۵=$\chi^{۲}$ است. جزئیات آماره والد در فصل هفتم بحث شده است.

۱۳-۳ پیش‌بینی با رگرسیون چندمتغیره

در فصل دوم مبحث پیش‌بینی را براساس رگرسیون دو متغیره بررسی کردیم. در اینجا نیز بحث مشابهی را برای رگرسیون K متغیره می‌توان مطرح کرد. بدین منظور معادله چندمتغیره را در نظر بگیرید آن را به‌صورت زیر می‌نویسیم:

$$Y_t = \beta_۱ + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + \cdots + \beta_K X_{Kt} + u_t$$

$$= [۱ \quad X_{۲t} \quad \cdots \quad X_{Kt}] \begin{bmatrix} \beta_۱ \\ \beta_۲ \\ \vdots \\ \beta_K \end{bmatrix} + u_t = \mathbf{x}'_t \boldsymbol{\beta} + u_t \qquad (۳-۷۷)$$

$$\mathbf{x}'_t = [۱ \quad X_{۲t} \quad \cdots \quad X_{Kt}]$$

مقدار پیش‌بینی Y برای سال f برابر است با:

$$Y_f = \mathbf{x}'_f \boldsymbol{\beta} + u_f \qquad (۳-۷۸)$$

مقدار تخمینی این پیش‌بینی برابر است با:

$$\hat{Y}_f = \mathbf{x}'_f \hat{\boldsymbol{\beta}} \qquad (۳-۷۹)$$

خطای پیش‌بینی برابر با تفاوت مقدار واقعی و پیش‌بینی Y در سال f می‌باشد:

$$e_f = Y_f - \hat{Y}_f = \mathbf{x}'_f \boldsymbol{\beta} + u_f - \mathbf{x}'_f \hat{\boldsymbol{\beta}} \qquad (۳-۸۰)$$

$$= -\mathbf{x}'_f (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + u_f$$

e_f یک متغیر تصادفی است که امید ریاضی آن برابر صفر است و واریانس آن برابر است با:

$$\begin{aligned}\text{var}(e_f) &= E(e_f^{۲}) = E[-\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta)+u_f][-\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta)+u_f]' \\ &= E[\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta)(\hat{\boldsymbol\beta}-\boldsymbol\beta)'\mathbf{x}_f]+E(u_f^{۲}) \\ &= \mathbf{x}'_f\sigma^{۲}(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{x}_f+\sigma^{۲} = \sigma^{۲}[۱+\mathbf{x}'_f(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{x}_f]\end{aligned} \quad (۳\text{-}۸۱)$$

توجه شود که $E[\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta)u_f]=\mathbf{x}'_fE(\hat{\boldsymbol\beta}-\boldsymbol\beta)E(u_f)=۰$ است زیرا $\hat{\boldsymbol\beta}$ تابعی از u_t است که $t<f$ می‌باشد و به دلیل عدم خودهمبستگی بین u_t و u_f، می‌توان نتیجه گرفت $\hat{\boldsymbol\beta}$ و u_f مستقل هستند.

با توجه به واریانس خطای پیش‌بینی می‌توان انحراف معیار خطای پیش‌بینی را حساب کرد که برابر با $SE(e_f)=\sqrt{\text{var}(e_f)}$ می‌باشد. مرز پیش‌بینی برابر با «به‌اضافه و منهای دو برابر انحراف معیار پیش‌بینی» می‌باشد. پیش‌بینی با مدل‌های چندمتغیره با استفاده از Eviews دقیقاً مشابه پیش‌بینی با رگرسیون یک متغیره است.

فاصله اطمینان پیش‌بینی برای مقدار پیش‌بینی (Y_f) عبارت است از:

$$t_{n-K}=\frac{e_f}{\sqrt{\text{var}(e_f)}} \Rightarrow t_{n-K}=\frac{Y_f-\hat{Y}_f}{\hat\sigma\sqrt{۱+\mathbf{x}'_f(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{x}_f}}$$

$$\Rightarrow\left(\hat{Y}_f-t_{۰/۰۵,n-K}\hat\sigma\sqrt{۱+\mathbf{x}'_f(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{x}_f}\le Y_f\le\hat{Y}_f+t_{۰/۰۵,n-K}\hat\sigma\sqrt{۱+\mathbf{x}'_f(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{x}_f}\right) \quad (۳\text{-}۸۲)$$

تعیین فاصله اطمینان پیش‌بینی برای امید ریاضی شرطی، یعنی $E(Y_t|X_t)=\mathbf{x}'_t\beta$، نیز مشابه بحث فوق می‌باشد. با توجه به اینکه برآورد امید ریاضی شرطی برابر با $\hat{Y}_t=\mathbf{x}'_t\hat\beta$ است لذا خطای پیش‌بینی برابر است با:

$$e_f^*=E(Y_f|X_f)-\hat{Y}_f=\mathbf{x}'_f\boldsymbol\beta-\mathbf{x}'_f\hat{\boldsymbol\beta}=-\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta) \quad (۳\text{-}۸۳)$$

امید ریاضی و واریانس e_f^* برابر است با:

$$\begin{aligned}E(e_f^*) &= ۰ \\ \text{var}(e_f^*) &= E(e_f^*e_f'^*)=E\left[\left(\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta)\right)\left(\mathbf{x}'_f(\hat{\boldsymbol\beta}-\boldsymbol\beta)\right)'\right] \\ &= \mathbf{x}'_fE\left[(\hat{\boldsymbol\beta}-\boldsymbol\beta)(\hat{\boldsymbol\beta}-\boldsymbol\beta)'\right]\mathbf{x}_f=\sigma^{۲}\mathbf{x}'_f(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{x}_f\end{aligned} \quad (۳\text{-}۸۴)$$

پیش‌بینی فاصله‌ای برای معادله رگرسیون یا امید ریاضی شرطی به صورت زیر به‌دست می‌آید:

$$t_{n-K} = \frac{e_f^*}{\sqrt{\text{var}(e_f^*)}} = \frac{E(Y_f \mid X_f) - \hat{Y}_f}{\hat{\sigma}\sqrt{\mathbf{x}'_f (\mathbf{X'X})^{-1} \mathbf{x}_f}}$$

$$\Rightarrow \left(\hat{Y}_f - t_{./۰۵, n-K} \hat{\sigma}\sqrt{\mathbf{x}'_f (\mathbf{X'X})^{-1} \mathbf{x}_f} \le E(Y_f \mid X_f) \le \hat{Y}_f + t_{./۰۵, n-K} \hat{\sigma}\sqrt{\mathbf{x}'_f (\mathbf{X'X})^{-1} \mathbf{x}_f} \right) \quad (۳-۸۵)$$

مسائل

۱- برای برآورد رگرسیون $Y_t = \beta_۰ + \beta_۱ X_{۱t} + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + u_t$ مقادیر زیر بر حسب انحراف از میانگین از نمونه‌ای به حجم $n = ۲۴$ حساب شده‌اند:

$$\sum y_i^۲ = ۶۰ \quad , \quad \sum x_{i۱}^۲ = ۱۰ \quad , \quad \sum x_{i۲}^۲ = ۳۰ \quad , \quad \sum x_{i۳}^۲ = ۲۰$$
$$\sum x_{i۱} y_i = ۷ \quad , \quad \sum x_{i۲} y_i = -۷ \quad , \quad \sum x_{i۳} y_i = -۲۶$$
$$\sum x_{i۱} x_{i۲} = ۱۰ \quad , \quad \sum x_{i۱} x_{i۳} = ۵ \quad , \quad \sum x_{i۲} x_{i۳} = ۱۵$$

الف) فرضیه $\beta_۱ + \beta_۲ + \beta_۳ = ۱$ را آزمون کنید.

ب) فرضیه $[\beta_۱ \ \ \beta_۲ \ \ \beta_۳] = [۱ \ \ ۱ \ \ ۲]$ را آزمون کنید.

ج) فرضیه‌های الف و ب چه تفاوتی دارند.

۲- مقادیر زیر از یک نمونهٔ ۲۰تایی به‌دست آمده است.

$$\sum Y_i = ۲۰ \quad , \quad \sum X_{i۱} = ۳۰ \quad , \quad \sum X_{i۲} = ۴۰$$
$$\sum Y_i^۲ = ۸۰ \quad , \quad \sum X_{i۱}^۲ = ۹۰ \quad , \quad \sum X_{i۲}^۲ = ۱۵۰$$
$$\sum X_{i۱} Y_i = ۶۰ \quad , \quad \sum X_{i۲} Y_i = ۹۰ \quad , \quad \sum X_{i۱} X_{i۲} = ۱۲۰$$

الف) مقادیر فوق را ابتدا بر حسب انحراف از میانگین به‌دست آورید.

ب) رگرسیون Y روی $X_۱$ و $X_۲$ را برآورد کنید و آماره t را برای هر یک از ضرایب حساب کنید.

ج) $R^۲$ و F را حساب کنید.

د) فرضیه برابری ضریب $X_۱$ و $X_۲$ را آزمون کنید.

۳- تابع تولید از نمونه‌ای به حجم n، به‌صورت زیر برآورد شده است:

$$\ln Q_i = ۱/۵ + ./۶۵۵ \ln K_i + ۰/۴۰۵ \ln L_i \qquad R^۲ = ./۹۵$$
$$(۱/۲) \quad (۳/۲) \qquad (۲/۵)$$

کوواریانس ضریب سرمایه و نیروی کار برابر با ۰/۰۵ است.

الف) آیا کشش تولیدی سرمایه و کار یکسان هستند؟

ب) آیا بازدهی ثابت نسبت به مقیاس وجود دارد؟

۳- ثابت کنید که $\hat{\mathbf{y}}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ است.

۴- ثابت کنید که R^2 مجذور ضریب همبستگی ساده بین $\mathbf{y}$ و $\hat{\mathbf{y}}$ است (که $\hat{\mathbf{y}}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ می‌باشد).

۵- فرق R^2 با $\bar{R}^2$ تعدیل شده چیست؟

۶- مدل رگرسیون زیر به‌صورت انحراف از میانگین می‌باشد:

$$y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

مقادیر زیر حساب شده‌اند:

$$\sum y_t^2 = 500 \qquad \sum x_{2t}^2 = 30 \qquad \sum x_{3t}^2 = 30 \qquad n = 20$$
$$\sum x_{2t} y_t = 30 \qquad \sum x_{3t} y_t = 20 \qquad \sum\sum x_{2t} x_{3t} = 0$$

الف) ضرایب β_2، β_3 و R^2 را برآورد کنید.

ب) فرضیه $\beta_3 = 7$ را آزمون کنید.

ج) فرضیه $\beta_2 = \beta_3 = 0$ را آزمون کنید.

د) فرضیه $\beta_2 = 3\beta_3$ را آزمون کنید.

۷- ثابت کنید که مجموع مجذورات پسماندهای (باقیمانده‌های) مقید بزرگتر از مجموع مجذورات غیرمقید است؟

۸- تابع تولید زیر را در نظر بگیرید:

$$Q_t = A L_{1t}^{\beta_1} L_{2t}^{\beta_2} K_t^{\beta_3} e^{u_t}$$

L_1 تعداد کارگران ماهر، L_2 تعداد کارگران غیرماهر، K سرمایه و u_t جمله اختلال است. در صورتی که همواره رابطه $L_1 + L_2 = 1000$ برقرار باشد آیا می‌توان تابع تولید را برآورد کرد؟

۹- در مسئله ۱ فصل اول، رگرسیون $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + u_i$ برآورد کنید.

۱۰- در مسئله ۳ اثر خالص و ناخالص X بر Y را حساب کنید.

۱۱- در مسئله ۱ فصل اول، ضرایب همبستگی جزئی را محاسبه و تحلیل کرده و با ضرایب همبستگی ساده مقایسه کنید.

۱۲- معادله زیر برآورد شده است:

$$\hat{Y}_i = ۲/۵ + ۰/۴۲X_{۲t} + ۱/۲X_{۳t} \qquad R^۲ = ۰/۷۵$$

آزمون معنی‌دار بودن رگرسیون را انجام دهید.

۱۳- مقادیر زیر برای یک رگرسیون دو متغیره حساب شده‌اند:

$$\mathbf{X'X} = \begin{bmatrix} ۳۳ & ۰ & ۰ \\ ۰ & ۴۰ & ۲۰ \\ ۰ & ۲۰ & ۶۰ \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} ۱۳۲ \\ ۲۴ \\ ۹۲ \end{bmatrix} \qquad \sum y_i^۲ = ۱۵۰$$

الف) حجم نمونه چقدر است؟

ب) ضرایب معادله رگرسیون را تخمین بزنید.

ج) $\hat{\sigma}^۲$ را حساب کنید.

د) $R^۲$ و $\overline{R}^۲$ را حساب کنید.

هـ) آماره t را برای هر یک از ضرایب حساب کنید.

و) آزمون معنی‌دار بودن رگرسیون را انجام دهید.

۱۴- در معادله $Y_i = \sum_{j=۱}^{K} \beta_j X_{ji} + u_i$ که شکل ماتریسی آن به‌صورت $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$ است، با نمونه‌ای به حجم n، تخمین $\hat{\boldsymbol{\beta}}$ به‌صورت زیر است:

$$\hat{\boldsymbol{\beta}}^n = (\mathbf{X'X})^{-۱}\mathbf{X'y}$$

حال اگر یک مشاهده به n مشاهدهٔ قبلی اضافه کنیم، ثابت کنید که تخمین جدید $\boldsymbol{\beta}$ که با $\hat{\boldsymbol{\beta}}^{n+۱}$ نشان می‌دهیم به‌صورت زیر است:

$$\hat{\boldsymbol{\beta}}^{n+۱} = \hat{\boldsymbol{\beta}}^n + \frac{(\mathbf{X'X})^{-۱}\mathbf{z}'_{n+۱}(Y_{n+۱} - \mathbf{z}_{n+۱}\hat{\boldsymbol{\beta}}^n)}{۱ + \mathbf{z}_{n+۱}(\mathbf{X'X})^{-۱}\mathbf{z}'_{n+۱}}$$

$$\mathbf{X} = \begin{bmatrix} ۱ & X_{۲۱} & \cdots & X_{K۱} \\ ۱ & X_{۲۲} & \cdots & X_{K۲} \\ \vdots & \vdots & & \vdots \\ ۱ & X_{۲n} & \cdots & X_{Kn} \end{bmatrix}$$

$$\mathbf{z}_{n+۱} = [۱ \quad X_{۲(n+۱)} \ \dots \ X_{K(n+۱)}]$$

۱۵- در مسئله ۱۴ ثابت کنید که رابطه زیر برقرار است:

$$RSS_{n+1} = RSS_n + \frac{(Y_{n+1} - \mathbf{z}_{n+1}\hat{\boldsymbol{\beta}}^n)}{1 + \mathbf{z}_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}'_{n+1}}$$

۱۶- در رگرسیون $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$، مراحل آزمون فرضیه $\beta_2 = \beta_4 = 0$ را توضیح داده و آماره F را تشکیل دهید.

۱۷- در معادله $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ کوواریانس $\hat{\beta}_2$ و $\hat{\beta}_3$ را حساب کنید. آیا امکان دارد که کوواریانس آنها برابر صفر شود.

۱۸- ثابت کنید که در رگرسیون $Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_K X_{Kt} + u_t$، تغییرات توضیح داده شده برابر است با:

$$ESS = \hat{\beta}_2 \sum x_{2t} y_t + \hat{\beta}_3 \sum x_{3t} y_t + \cdots + \hat{\beta}_K \sum x_{Kt} y_t$$

۱۹- معادله $Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_K X_{Kt} + u_t$ را به‌صورت $y_t = \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + u_t$ بنویسید. حروف کوچک بیانگر انحراف از میانگین هستند.

۲۰- رگرسیون‌های زیر برآورد شده‌اند:

$$\hat{Y}_t = \hat{\alpha}_1 + 0/95 X_{2t}$$
$$\hat{Y}_t = \hat{\alpha}_2 + 0/85 X_{3t}$$
$$\hat{X}_{3t} = \hat{\alpha}_3 + 0/75 X_{2t}$$
$$\hat{X}_{2t} = \hat{\alpha}_4 + 0/65 X_{3t}$$

بر اساس اطلاعات فوق، ضرایب β_2 و β_3 را در معادله $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ برآورد کنید.

۲۱- رگرسیون $Y_t = \alpha + \beta Y_t^* + u_t$ را داریم که Y_t مقدار واقعی و Y_t^* مقدار پیش‌بینی می‌باشد. فرضیه انتظارات عقلایی ادعا می‌کند که شکل‌گیری انتظارات به‌طور متوسط درست است، زیرا خطای پیش‌بینی دارای میانگین صفر می‌باشد ($E(Y_t) = E(Y_t^*)$ است). نتایج حاصل از برآورد مدل فوق عبارت است از:

$$Y_t = 0/25 + 0/95 Y_t^* \qquad RSS = 25/4 \ , \ n = 32$$
$$(0/2) \quad (6/6)$$

اگر $\overline{Y}^* = 10$ ، $\sum (Y_t^* - \overline{Y}^*)^2 = 40/85$ باشد. آیا می‌توان نتایج فوق را مطابق با فرضیه انتظارات عقلایی دانست؟

۲۲- در رگرسیون $Y_t=\beta_1+\beta_2X_{2t}+\cdots+\beta_KX_{Kt}+u_t$ آیا درست است که آزمون فرضیه $H_0:\beta_2=\cdots=\beta_K=0$ را معادل به فرضیه $H_0:R^2=0$ بدانیم؟ چرا؟

۲۳- اگر در گرسیون $Y_t^*=\beta_1+\beta_2X_{2t}^*+\beta_3X_{3t}^*+u_t$ متغیرها استاندارد شده باشند، ثابت کنید که روابط زیر برقرار است (r_{ij} ضریب همبستگی بین متغیرها را نشان می‌دهد: ۱ برای Y، ۲ برای X_2 و ۳ برای X_3)

الف) $\mathbf{X'X}=n\begin{bmatrix}1 & r_{23}\\ r_{23} & 1\end{bmatrix}$

ب) $\mathbf{X'y}=n\begin{bmatrix}r_{12}\\ r_{13}\end{bmatrix}$

ج) $\hat{\beta}_1=0$

د) $\hat{\beta}_2=\dfrac{r_{12}-r_{23}r_{13}}{1-r_{23}^2}$, $\hat{\beta}_3=\dfrac{r_{13}-r_{23}r_{12}}{1-r_{23}^2}$

۲۴- ثابت کنید که اگر به جای معادله $Y_t=\alpha_1+\alpha_2X_{2t}+u_t$ متغیر X_{3t} را اضافه کنیم، اولاً R^2 افزایش می‌یابد و ثانیاً در صورتی $\bar{R}^2$ افزایش می‌یابد که آماره t برای ضریب X_{3t} بزرگتر از ۱ باشد.

۲۵- رگرسیون $Y_t=\beta_1+\beta_2X_{2t}+u_t$ را داریم. باقیمانده‌های این مدل را با e_t و مقادیر برآوردی را با $\hat{Y}_t$ نشان می‌دهیم.

الف) در رگرسیون $e_t=\alpha_1+\alpha_2X_{2t}+\varepsilon_t$ ضریب تعیین (R^2) چه چیزی را بیان می‌کند. ثابت کنید که nR^2 برابر با آماره ضریب لاگرانژ است.

ب) R^2 حاصل از رگرسیون $Y_t=\gamma_1+\gamma_2\hat{Y}_t+v_t$ چه چیزی را نشان می‌دهد و چه رابطه‌ای با R^2 در معادله $Y_t=\beta_1+\beta_2X_{2t}+u_t$ دارد.

ج) در رگرسیون $Y_t=a_1+a_2e_t+w_t$ ضرایب a_1 و a_2 را برآورد کنید. R^2 این معادله چه تفسیری دارد؟

۲۶- رگرسیون $Y_t=\beta_1+\beta_2X_{2t}+\beta_3X_{3t}+u_t$ را در نظر بگیرید. حال رگرسیون‌های زیر نیز برآورد شده‌اند:

$$X_{2t} = \alpha_1 + \alpha_2 X_{3t} + v_{2t}$$
$$X_{3t} = \gamma_1 + \gamma_2 X_{2t} + v_{3t}$$

با استفاده از باقیمانده‌های معادلات فوق، معادله زیر را برآورد می‌کنیم:

$$\hat{v}_{3t} = b_1 + b_2 \hat{v}_{2t} + \varepsilon_t$$

الف) ثابت کنید که $\hat{b}_2 = \hat{\beta}_2$ است.

ب) اگر رگرسیون $Y_t = \gamma_1 + \gamma_2 \hat{v}_{2t} + \gamma_3 \hat{v}_{3t} + u_t$ را برآورد کنیم، رابطه بین $\hat{\gamma}_2$ و $\hat{\beta}_2$ را به دست آورید.

ضمیمه فصل ۳: رگرسیون چند متغیره در Stata

برآورد معادلات رگرسیون چند متغیره در Stata

فایل data3

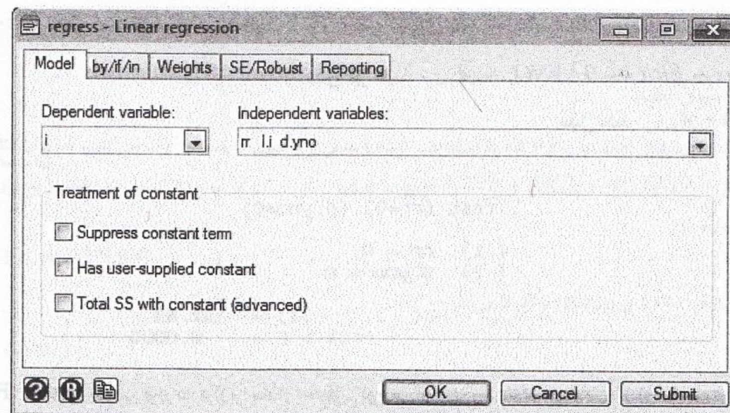
برآورد رگرسیون چندمتغیره، مشابه رگرسیون ساده است. بدین منظور معادله زیر را در نظر بگیرید:

$$I_t = \alpha + \beta RR_t + \theta \Delta YNO_t + \gamma I_{t-1}$$

در اینجا برای وارد کردن تفاضل yno از d.yno و برای وقفه I_{t-1} از l.i استفاده می‌شود. بنابراین با انتخاب مسیر زیر مشابه رگرسیون ساده عمل می‌کنیم:

Linear regression → Linear models and related → Statistics

متغیرها را به صورت زیر وارد می‌کنیم:



نتایج به‌صورت زیر به دست می‌آید که کاملاً مشابه رگرسیون ساده است:

```
. - preserve
regress i rr   l.i  d.yno

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  3,    31) =  184.73
       Model |  4.1221e+10     3  1.3740e+10           Prob > F      =  0.0000
    Residual |  2.3058e+09    31  74380117.7           R-squared     =  0.9470
-------------+------------------------------           Adj R-squared =  0.9419
       Total |  4.3527e+10    34  1.2802e+09           Root MSE      =  8624.4

------------------------------------------------------------------------------
           i |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |  -246.8273   200.1467    -1.23   0.227    -655.0293    161.3746
           i |
         L1. |   .8194636   .0523257    15.66   0.000     .7127447    .9261825
         yno |
         D1. |    1.18792   .1583372     7.50   0.000     .8649894    1.510851
       _cons |   5742.107    5095.64     1.13   0.268     -4650.52    16134.73
------------------------------------------------------------------------------
.
```

ضریب RR_t (rr)
ضریب I_{t-1} (L1.)
ضریب ΔYNO_t (D1.)
عرض از مبدأ (_cons)

آزمون محدودیت‌ها (آزمون والد) در Stata

الف) آزمون محدودیت‌های صفری:

می‌خواهیم به‌طور همزمان صفربودن چند ضریب را آزمون کنیم. مثلاً صفر بودن ضریب rr و ΔYNo.

بدین منظور، بعد از برآورد معادله مورد نظر (در اینجا از معادله $I_t = \alpha + \beta RR_t + \theta \Delta YNO_t + \gamma I_{t-1}$ استفاده کرده‌ایم)، فرمان زیر را اجرا می‌کنیم:

test (rr=0) (l.yno=0)

این فرضیه معادل با $\beta = \theta = 0$ است که می‌توان از آزمون F برای مقایسه دو مدل مقید و غیرمقید استفاده نمود:

$I_t = \alpha + \beta RR_t + \theta \Delta YNO_t + \gamma I_{t-1}$ مدل غیر مقید

$I_t = \alpha + \gamma I_{t-1}$ مدل مقید

نتیجه آزمون عبارت است از:

```
. test (rr=0) (d.yno=0)

 ( 1)  rr = 0
 ( 2)  D.yno = 0

       F(  2,    31) =   31.83
            Prob > F =    0.0000
```

چون مقدار F=۳۱/۸۳ در ناحیه بحرانی قرار دارد (زیرا مقدار احتمال کوچکتر از ۰/۵۰ است)، لذا فرضیه H_0 (وجود محدودیت‌ها) رد می‌شود.

ب) آزمون محدودیت‌های خطی:

به عنوان مثال، محدودیت $\theta - 2\gamma = 0$ و $0/01\beta + \theta = 120$ را درنظر بگیرید.

ابتدا معادله مورد نظر را تخمین می‌زنیم و سپس فرمان زیر را اجرا می‌کنیم:

```
test (d.yno-2*l.i==0) (0.01*rr+d.yno==120)
```

نتیجه آزمون عبارت است از:

```
. test (d.yno-2*l.i==0) (0.01*rr+d.yno==120)

 ( 1)  D.yno - 2 L.i = 0
 ( 2)  .01 rr + D.yno = 120

       F(  2,    31) = 2082.10
            Prob > F =    0.0000
```

چون مقدار F=۲۰۸۲/۱ در ناحیه بحرانی قرار دارد لذا فرضیه $H_\cdot$ مبنی بر برقرار بودن این محدودیت‌ها، رد می‌شود.

ج) آزمون محدودیت‌های غیرخطی

این آزمون نیز مشابه آزمون محدودیت‌های خطی است با این تفاوت که بایستی آن را به شکل دیگری وارد نماییم. به‌عنوان مثال مدل $Y_t = \alpha + \beta X_t + \beta\gamma X_{t-1} + \gamma Y_{t-1} + u_t$ را در نظر بگیرید. این مدل ضریب X_{t-1} برابر با حاصل‌ضرب ضرایب X_t و Y_{t-1} است. برای انجام این آزمون ابتدا معادله را به صورت زیر برآورد می کنیم:

```
reg y x l.x l.y
```

حال قید مورد نظر را با فرمان زیر آزمون می‌کنیم:

```
testnl _b[x]*_b[l.y]=_b[l.x]
```

بر اساس مقدار F در خصوص معنی‌دار بودن قیدها تصمیم‌گیری می کنیم.

تخمین رگرسیون مقید

برای برآورد رگرسیون مقید، ابتدا هر یک از قیدها را تعریف کرده و سپس رگرسیون مقید را برآورد می‌کنیم. به‌عنوان مثال می‌خواهیم رگرسیون $I_t = \alpha + \beta RR_t + \theta \Delta YNO_t + \gamma I_{t-1}$ با توجه به قید $\theta - 2\gamma = 0$ برآورد کنیم. ابتدا قید را به‌صورت زیر تعریف می‌کنیم (قید شماره ۱):

```
constraint 1 d.yno-2*l.i=0
```

حال رگرسیون مقید را با فرمان cnsreg برآورد می‌کنیم:

```
cnsreg i rr d.yno l.i, c(1)
```

نتیجهٔ تخمین مقید عبارت است از:

```
. cnsreg i rr d.yno l.i, c(1)

Constrained linear regression                     Number of obs =      35
                                                  F(  2,     32) =  248.68
                                                  Prob > F       =  0.0000
                                                  Root MSE       =  9067.8
 ( 1)  D.yno - 2 L.i = 0
```

i	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rr	-355.4319	203.2331	-1.75	0.090	-769.4042	58.54047
yno D1.	1.481907	.0766643	19.33	0.000	1.325747	1.638067
i L1.	.7409533	.0383321	19.33	0.000	.6628733	.8190333
_cons	8901.894	5116.793	1.74	0.092	-1520.672	19324.46

اگر دو قید به صورت $\theta - 2\gamma = 0$ و $0/01\beta + \theta = 25$ داشته باشیم، ابتدا هر یک از قیدها را جداگانه با شماره‌های ۱ و ۲ به‌صورت زیر تعریف می‌کنیم:

```
constraint 1  d.yno-2*l.i=0
constraint 2  0.01*rr+d.yno=25
```

حال رگرسیون مقید را با فرمان زیر برآورد می‌کنیم:

```
cnsreg  i  rr  d.yno  l.i,  c(1 2)
```



```
. cnsreg i rr d.yno l.i, c(1 2)

Constrained linear regression
                                                  Number of obs =      35
 ( 1)  D.yno - 2 L.i = 0                          Root MSE       =   23257
 ( 2)  .01 rr + D.yno = 25
```

i	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rr	2407.991	16.56588	145.36	0.000	2374.288	2441.695
yno D1.	.920087	.1656588	5.55	0.000	.5830516	1.257122
i L1.	.4600435	.0828294	5.55	0.000	.2915258	.6285611
_cons	55063.65	9822.637	5.61	0.000	35079.34	75047.96

فصل چهارم

نقض فروض کلاسیک

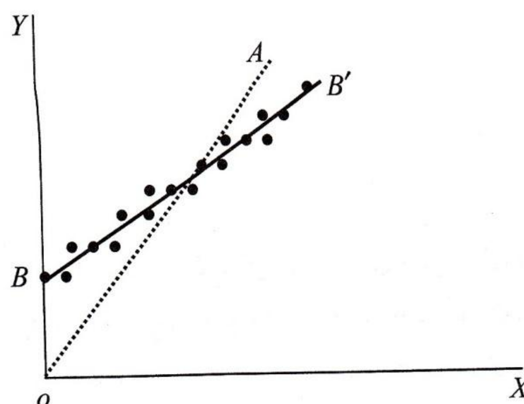
۴-۱ مقدمه

در فصل دوم فروض کلاسیک را مطرح کردیم که مبنای اصلی برآورد معادلات رگرسیون است. همان‌طور که دیدیم تمام نتایج و تحلیل‌های مربوط به تخمین معادلات، عمدتاً بر مبنای این فروض قرار دارد. در این راستا با چند مسئله مواجه‌ایم. اولاً چگونه می‌توان نقض فروض کلاسیک را تشخیص داد. ثانیاً در عمل چه چیزی باعث نقض فروض کلاسیک می‌شود. ثالثاً وقتی بدون توجه به نقض فروض کلاسیک، به تخمین معادلات و تفسیر نتایج می‌پردازیم با چه مسائل و مشکلاتی مواجه می‌شویم.

۴-۲ فرض اول: صفر بودن میانگین خطا

اولین فرض کلاسیک این است که امید ریاضی u_t برابر با صفر است. از آنجا که $e_t = Y_t - \hat{Y}_t$ است، لذا وقتی میانگین خطا صفر باشد بدان معنا است که مقدار تخمینی به‌طور متوسط برابر با مقدار واقعی است و در نتیجه، خطای متوسط یا امید ریاضی خطا برابر صفر خواهد بود. می‌توان ثابت کرد که اگر $E(u_t)$ برابر با مقدار ثابتی باشد، آنگاه شیب معادله رگرسیون بدون تورش، ولی برآورد عرض از مبدأ باتورش خواهد بود. اگر $E(u_t)$ ثابت نباشد، آنگاه برآورد شیب معادله رگرسیون نیز دچار تورش خواهد شد.

از طرف دیگـر اگـر معادلـه رگرسـیون دارای عـرض از مبـدأ باشـد، فـرض $E(u_t)=0$ نقـض نمی‌شود. اما اگر رگرسیون فاقد عرض از مبدأ باشد، مقدار متوسط خطاها الزاماً صفر نخواهد شد و دارای آثار نامطلوبی خواهد بود؛ اولاً R^2 که برابر با نسبت ESS/TSS اسـت ممکـن اسـت منفـی شود. این موضوع بدین معنی است که متوسط نمونه ($\overline{Y}$) می‌تواند تغییرات Y را بهتر از متغیرهای توضیحی، توضیح دهد. به عبـارت دیگـر $\overline{Y}$ بـه‌عنـوان بـرآوردی از Y، بهتـر از معادلـه رگرسـیون می‌باشد.[1] ثانیاً، وقتی معادله رگرسیون فاقد عرض از مبدأ باشـد مـی‌توانـد بـه‌طـور بـالقوه منجـر بـه تورش شدید در تخمین شیبِ معادله شود. بدین منظور نمودار(۱-۴) را در نظر بگیرید.



نمودار ۱-٤: حذف عرض از مبدأ و تأثیر آن بر شیب

خط OA بدون عرض از مبدأ و خط BB' با عـرض از مبـدأ مـی‌باشـد. تـصور کنیـد کـه خـط رگرسیون صحیح توسط خط BB' توصیف شود. در این صورت اگـر جملـه ثابـت حـذف شـود، تخمین شیب که با OA نشان داده شده است، دارای تورش خواهد بود.

در رگرسیون فاقد عرض از مبـدأ، خطـا برابـر بـا $e_t=Y_t-\hat{Y}_t=Y_t-\hat{\beta}X_t$ اسـت. از حـداقل شدن مجموع مجذور خطاها خواهیم داشت:

$$\sum e_t^2=\sum(Y_t-\hat{\beta}X_t)^2 \qquad (۴-۱)$$

---

[1] توجه شود که در رگرسیون $Y_t=\alpha+u_t$ برآورد α برابر با $\overline{Y}$ است. در این معادله، $R^2=0$ است. حـال اگـر در رگرسیون $Y_t=\beta X_t+u_t$ مقدار R^2 منفی باشد، بدان معنا است که $\hat{Y}_t=\hat{\alpha}=\overline{Y}$ بهتر از $\hat{Y}_t=\hat{\beta}X_t$ می‌باشد.

$$\frac{\partial e_t^{۲}}{\partial \hat{\beta}} = -۲\sum(Y_t - \hat{\beta}X_t)X_t = ۰ \Rightarrow \sum(Y_t - \hat{\beta}X_t)X_t = ۰ \Rightarrow \sum e_t X_t = ۰ \qquad (۴-۲)$$

با حل (۲-۴) تخمین β به دست می‌آید:

$$\hat{\beta}^* = \frac{\sum X_t Y_t}{\sum X_t^{۲}} \qquad (۴-۳)$$

لذا در اینجا از شرط $\sum e_t X_t = ۰$ استفاده شده است، اما شرط $\sum e_t = ۰$ که برای تخمین α به‌کار می‌رفت، مورد استفاده قرار نمی‌گیرد. لذا ممکن است شرط $\sum e_t = ۰$ برقرار نباشد و بدان معنا است که میانگین خطاها، صفر نخواهد شد. اگر شرط $\sum e_t = ۰$ را به‌کار ببریم، آنگاه تخمین β به‌صورت زیر خواهد شد:

$$\sum e_t = ۰ \ \Rightarrow \ \sum Y_t - \hat{\beta}^{**}\sum X_t = ۰ \ \Rightarrow \ \overline{Y} = \hat{\beta}^{**}\overline{X} \ \Rightarrow \hat{\beta}^{**} = \frac{\overline{Y}}{\overline{X}} \qquad (۴-۴)$$

بنابراین، تخمین دیگری برای شیب معادله رگرسیون به دست می آید که با $\hat{\beta}^*$ متفاوت است. بدین معنا که $\hat{\beta}^*$ مجموع مجذور خطاها یعنی $\sum e_t^{۲} = \sum(Y_t - \hat{\beta}X_t^{۲})$ را حداقل می‌کند که معادل با شرط $\sum e_t X_t = ۰$ است، در حالی که $\hat{\beta}^{**}$ مجموع خطاها را صفر می‌کند که معادل با شرط $\sum e_t = ۰$ است.

می‌توان نشان داد که وقتی عرض از مبدأ را از مدل صحیح یعنی $Y_t = \alpha + \beta X_t + u_t$ حذف می‌کنیم و به‌جای آن از مدل $Y_t = \beta X_t + v_t$ استفاده می‌کنیم، باعث می شود که تخمین شیب (یعنی $\hat{\beta}^*$) دچار تورش شود:

$$\hat{\beta}^* = \frac{\sum X_t Y_t}{\sum X_t^{۲}} = \frac{\sum X_t(\alpha + \beta X_t + u_t)}{\sum X_t^{۲}} = \alpha\frac{\sum X_t}{\sum X_t^{۲}} + \beta + \frac{\sum X_t u_t}{\sum X_t^{۲}}$$

$$E(\hat{\beta}^*) = \alpha\frac{\sum X_t}{\sum X_t^{۲}} + \beta + \frac{\sum X_t E(u_t)}{\sum X_t^{۲}} = \beta + \alpha\frac{\sum X_t}{\sum X_t^{۲}} = \beta + \alpha\frac{n\overline{X}}{\sum X_t^{۲}} \qquad (۴-۵)$$

توجه شود که در مدل صحیح، $E(u_t) = ۰$ است در حالی که در مدلی که عرض از مبدأ آن را حذف کرده‌ایم، امید ریاضی جملهٔ خطا صفر نمی‌باشد ($E(v_t) \neq ۰$).

بنابراین، $\hat{\beta}^*$ دارای تورش است که علامت تورش عمدتاً بستگی به علامت α و همچنین بستگی به علامت $\overline{X}$ دارد. تورش $\hat{\beta}^*$ وقتی صفر می‌شود که α یا $\overline{X}$ صفر باشد. یعنی یا واقعاً عرض از

مبدأ صفر باشد و یا متغیر X دارای مقادیر مثبت و منفی باشد به‌گونه‌ای که میانگین آن صفر گردد.

از طرف دیگر وقتی شرط $\sum e_t = 0$ برقرار باشد، آنگاه $\overline{\hat{Y}} = \overline{Y}$ است، زیرا $Y_t = \hat{Y}_t + e_t$ است و لذا خواهیم داشت:

$$\sum Y_t = \sum \hat{Y}_t + \sum e_t \Rightarrow \overline{Y} = \overline{\hat{Y}} + \overline{e} \Rightarrow \overline{Y} = \overline{\hat{Y}} \quad , \quad \overline{e} = 0 \qquad (۴-۶)$$

برابری $\overline{\hat{Y}} = \overline{Y}$ بدان معنا است که معادله رگرسیون، لااقل مقادیر متوسط Y را دقیقاً به ما خواهد داد در حالی که معادلهٔ بدون عرض از مبدأ، چنین تضمینی را نمی‌دهد. زیرا $\overline{e} \neq 0$ بوده و لذا $\overline{Y} \neq \overline{\hat{Y}}$ می‌باشد.

نکته دیگر این است که وقتی α را حذف کنیم، ممکن است R^2 منفی شود. طبق تعریف، R^2 برابر است با:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

بدین منظور، RSS را در دو حالت بررسی می کنیم:

۱- اگر معادله شامل عرض از مبدأ باشد، همان‌طور که دیدیم $TSS = \sum y_t^2$ و $ESS = \sum \hat{y}_t^2 = \hat{\beta}^2 \sum x_t^2$ است. لذا $ESS \geq 0$ است و در نتیجه $RSS \leq TSS$ بوده که موجب می‌شود تا $R^2 \geq 0$ باشد. از طرف دیگر چون $\sum e_t = 0$ است لذا $\overline{Y} = \overline{\hat{Y}}$ خواهد بود.

۲- اگر معادله فاقد عرض از مبدأ باشد، در این صورت $Y_t = \beta X_t + u_t$ و $\hat{Y}_t = \hat{\beta} X_t$ است که در این حالت، شرط $\sum e_i = 0$ برقرار نیست. لذا میانگین Y_i و $\hat{Y}_i$ برابر نیستند.

$$\sum e_i \neq 0 \Rightarrow \sum (Y_i - \hat{Y}_i) \neq 0 \Rightarrow \sum Y_i - \sum \hat{Y}_i \neq 0 \Rightarrow \overline{Y} \neq \overline{\hat{Y}}$$

از طرف دیگر TSS برابر است با:

$$TSS = \sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i + e_i - \overline{Y})^2 = \sum [(\hat{Y}_i - \overline{Y}) + e_i]^2$$
$$= \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2 + 2\sum e_i (\hat{Y}_i - \overline{Y})$$

جملهٔ آخر برابر است با:

$$2\sum e_i (\hat{Y}_i - \overline{Y}) = 2\sum e_i \hat{Y}_i - 2\overline{Y} \sum e_i = 2\overline{Y}\,\overline{e}$$

زیرا $\sum e_i\hat{Y}_i = 0$ است. از طرف دیگر در رگرسیون فاقد عرض از مبدأ، $\bar{e} \neq 0$ است که با توجه به $\bar{Y} = \bar{\hat{Y}} + \bar{e}$، خواهیم داشت:

$$2\bar{Y}\bar{e} = 2\bar{Y}(\bar{Y} - \bar{\hat{Y}})$$

بنابراین، TSS برابر است با:

$$TSS = \sum(\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2\bar{Y}(\bar{Y} - \bar{\hat{Y}})$$

دیدیم که $\sum(\hat{Y}_i - \bar{Y})^2$ معادل با ESS است که در اینجا برابر است با:

$$ESS = TSS - RSS - 2\bar{Y}(\bar{Y} - \bar{\hat{Y}})$$

در رگرسیون با عرض از مبدأ، $\bar{Y} = \bar{\hat{Y}}$ است و لذا $ESS=TSS-RSS$ می‌باشد، ولی در رگرسیون فاقد عرض از مبدأ، $\bar{Y} \neq \bar{\hat{Y}}$ است و اگر $\bar{Y} > \bar{\hat{Y}}$ باشد، آنگاه ممکن است $ESS = (TSS - RSS) - 2\bar{Y}(\bar{Y} - \bar{\hat{Y}})$ منفی شود. در این صورت، ESS منفی شده و موجب منفی‌شدن R^2 خواهد شد.

۴-۳ فرض دوم: واریانس همسانی

اگر جمله خطا دارای واریانس ثابت نباشد، آن را «واریانس ناهمسانی»[1] می‌گویند که در مقابل «واریانس همسانی»[2] قرار دارد. در اینجا به بررسی ماهیت واریانس ناهمسانی، پیامدها، آزمون‌های تشخیصی و برطرف کردن آن می‌پردازیم.

۴-۳-۱ ماهیت واریانس ناهمسانی

واریانس ناهمسانی یکی از مشکلاتی است که در برآورد معادلات رگرسیون با آن مواجه می‌شویم. واریانس ناهمسانی بدان معنا است که واریانس u_t (و به‌دنبال آن واریانس Y_t) ثابت نیست و معمولاً همراه با یک یا چند متغیر، افزایش می‌یابد. به‌ویژه، این پدیده در داده‌های مقطعی بسیار به چشم می‌خورد. به‌عنوان مثال وقتی مصرف یک کالا را (Y) را روی درآمد (X) برازش

---

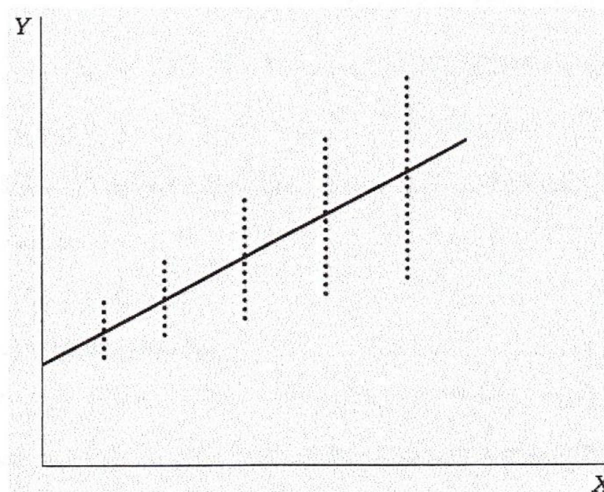
[1] heteroskedasticity
[2] hemoskedasticity

می‌کنیم و از داده‌های مقطعی (مانند گروه‌های درآمدی) استفاده می‌کنیم، در گروه‌های درآمدی پایین، مقدار مصرف کالای مورد نظر به‌هم نزدیک است ولی در گروه‌های درآمدی بالا، تفاوت‌های بیشتری وجود دارد. این موضوع باعث واریانس ناهمسانی می‌شود که طبق آن، σ_t^2 تابعی از درآمد (X_t) خواهد بود. در داده‌های سری زمانی نیز ممکن است نوسانات یک متغیر مانند بازده سهام، نرخ ارز و نرخ تورم در طول زمان افزایش یابد که شواهدی از واریانس ناهمسانی هستند.

در رگرسیون $Y_t = \alpha + \beta X_t + u_t$ واریانس u_t و واریانس شرطی Y_t برابر است با:

$$\sigma_t^2 = E(u_t^2 \mid X_t) = \text{var}(Y_t \mid X_t) \qquad (۴-۸)$$

نمودار زیر بیانگر واریانس ناهمسانی است:



نمودار ۲-٤: واریانس ناهمسانی

نمودار ۲-۴ نشان می‌دهد که با افزایش X، پراکندگی Y (واریانس Y) حول میانگین شرطی (خط رگرسیون) در حال افزایش است. بنابراین، واریانس شرطی Y و u تابعی از X است.

$$\sigma_t^2 = E(u_t^2 \mid X_t) = \text{var}(Y_t \mid X_t) = f(X_t) \qquad (۴-۹)$$

از طرف دیگر، می‌توان رابطه فوق را برای u_t^2 به‌صورت زیر نوشت:

$$u_t^2 = E(u_t^2 \mid X_t) + \varepsilon_t = \sigma_t^2 + \varepsilon_t = f(X_t) + \varepsilon_t \quad ; \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \qquad (۴-۱۰)$$

از آنجا که مقادیر u_t را نداریم از برآورد آنها، یعنی e_t، استفاده می‌کنیم:

$$e_t^2 = f(X_t) + \varepsilon_t \qquad (۴-۱۱)$$

معادله فوق یک معادله رگرسیون است که e_t^2 را روی X_t برازش می‌کند. تابع f نوع رابطه بین e_t^2 و X_t را مشخص می‌کند که می‌تواند هر شکلی داشته باشد. معمولاً هر یک از روش‌های تشخیص واریانس ناهمسانی شکل خاصی از تابع f را در نظر می‌گیرند. البته، نوع دیگری از واریانس ناهمسانی وجود دارد که طبق آن، واریانس u_t تابعی از خطاهای گذشته (u_{t-i}ها) است. این نوع مدل‌ها معروف به مدل‌های ARCH[1] می‌باشند که در فصل چهاردهم بحث خواهند شد.

۲-۳-٤ پیامدهای واریانس ناهمسانی

۱- واریانس ناهمسانی، ویژگی بدون تورش بودن تخمین‌زننده‌های OLS را تغییر نمی‌دهد.

۲- واریانس ناهمسانی، ویژگی سازگاری تخمین‌زننده‌های OLS را تغییر نمی‌دهد.

۳- واریانس ناهمسانی، موجب می‌شود که تخمین‌زننده‌های OLS، کارا نباشند، یعنی حداقل واریانس را ندارند. به‌طور کلی برآورد واریانس جملهٔ خطا دچار تورش می‌شود.

۴- از آنجا که واریانس‌ها به‌درستی برآورد نمی‌شوند، لذا آزمون فرضیه‌هایی که با t، F و χ^2 انجام می‌شوند، قابل اطمینان نخواهند بود.

۳-۳-٤ برآورد واریانس مستحکم[2] وایت

یکی از پیامدهای واریانس ناهمسانی آن است که برآورد $\hat{\sigma}^2$ دارای تورش است و لذا تمامی آزمون‌های فرضیه را مخدوش می‌کند. برای حل این مشکل و برآورد بهتری از واریانس، روشی توسط وایت معرفی شده است که در اینجا آن را تشریح می‌کنیم.

در فصل دوم برای معادله رگرسیون سادهٔ $Y_t = \alpha + \beta X_t + u_t$، تخمین β از روش OLS را به‌صورت زیر به‌دست آوردیم:

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \sum w_t Y_t = \beta + \sum w_t u_t \quad ; \quad w_t = \frac{x_t}{\sum x_t^2}$$

---

[1] Autoregressive Conditional Heteroskedasticity
[2] robust

دیدیم که واریانس $\hat{\beta}$ در صورت واریانس همسانی، برابر است با:

$$\operatorname{var}(\hat{\beta}) = E(\hat{\beta}-\beta)^{2} = \sum w_t^{2}E(u_t^{2}) = \frac{\sum x_t^{2}E(u_t^{2})}{(\sum x_t^{2})^{2}} = \frac{(\sum x_t^{2})\sigma^{2}}{(\sum x_t^{2})^{2}} = \frac{\sigma^{2}}{\sum x_t^{2}}$$

حال در صورت واریانس ناهمسانی، واریانس $\hat{\beta}$ برابر است با:

$$\operatorname{var}(\hat{\beta}) = \frac{\sum x_t^{2}E(u_t^{2})}{(\sum x_t^{2})^{2}} = \frac{\sum x_t^{2}\sigma_t^{2}}{(\sum x_t^{2})^{2}} \qquad (۴-۱۲)$$

بنابراین در صورت واریانس همسانی، واریانس $\hat{\beta}$ برابر با $\sigma^{2}/\sum x_t^{2}$ است که $\hat{\sigma}^{2} = \frac{\sum e_t^{2}}{n-2}$ می‌باشد. لذا واریانس $\hat{\beta}$ به‌صورت زیر محاسبه می‌شود:

$$\operatorname{var}(\hat{\beta}) = \frac{\hat{\sigma}^{2}}{\sum x_t^{2}} = \frac{\sum \frac{1}{n-2}e_t^{2}}{\sum x_t^{2}} \qquad (۴-۱۳)$$

معادله فوق نشان می‌دهد که در محاسبه واریانس $\hat{\beta}$ وزن e_t^{2}ها یکسان بوده و برابر با $\frac{1}{n-2}$ می‌باشد.

وایت (۱۹۸۰) پیشنهاد می‌کند که برای برآورد سازگار $\operatorname{var}(\hat{\beta})$ از (۱۲-۴) ه صورت زیر استفاده کنیم که مشابه (۱۳-۴) است ولی به‌جای $\frac{1}{n-2}$ از $\frac{x_t^{2}}{\sum x_t^{2}}$ استفاده می‌شود:

$$\operatorname{var}(\hat{\beta})_{White} = \frac{\sum x_t^{2}e_t^{2}}{(\sum x_t^{2})^{2}} = \frac{\sum \frac{x_t^{2}e_t^{2}}{\sum x_t^{2}}}{\sum x_t^{2}} = \frac{\sum k_t e_t^{2}}{\sum x_t^{2}} \quad ; \quad k_t = \frac{x_t^{2}}{\sum x_t^{2}} \qquad (۴-۱۴)$$

توجه شود که در اینجا تخمین σ_t^{2} را نمی‌توانیم به‌دست آوریم، زیرا e_t^{2} فقط یک مشاهده است. لذا فقط تخمین سازگار واریانس $\hat{\beta}$ را به‌دست آورده‌ایم، یعنی تخمین سازگار برای واریانس تخمین‌زننده‌های OLS. بنابراین روش وایت از باقیمانده‌های OLS استفاده می‌کند و واریانس $\hat{\beta}$ را اصلاح می‌کند به‌گونه‌ای که سازگار شود. لذا می‌توان آن را برای آزمون‌های فرضیه مورد استفاده قرار داد.

حال اگر رگرسیون K متغیره باشد، واریانس $\hat{\beta}_k$ که ضریب X_{kt} است به‌صورت زیر پیشنهاد می‌شود:

$$\text{var}(\hat{\beta}_k)_{White} = \frac{\sum_{t=1}^{n} \hat{\varepsilon}_{kt}^2 e_t^2}{\left(\sum_{t=1}^{n} \hat{\varepsilon}_{kt}^2\right)^2} \quad (۴-۱۵)$$

که ε_{kt} جملات خطای معادله رگرسیون $X_{kt} = \alpha_0 + \alpha_1 e_t + \varepsilon_{kt}$ است.

برآورد واریانس وایت در Eviews فایل data5

به منظور برآورد واریانس وایت، با استفاده از مسیر Quick → Estimate Equation پنجره زیر را باز می‌کنیم:

Equation Estimation

Specification | Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.

Y C X2 X3

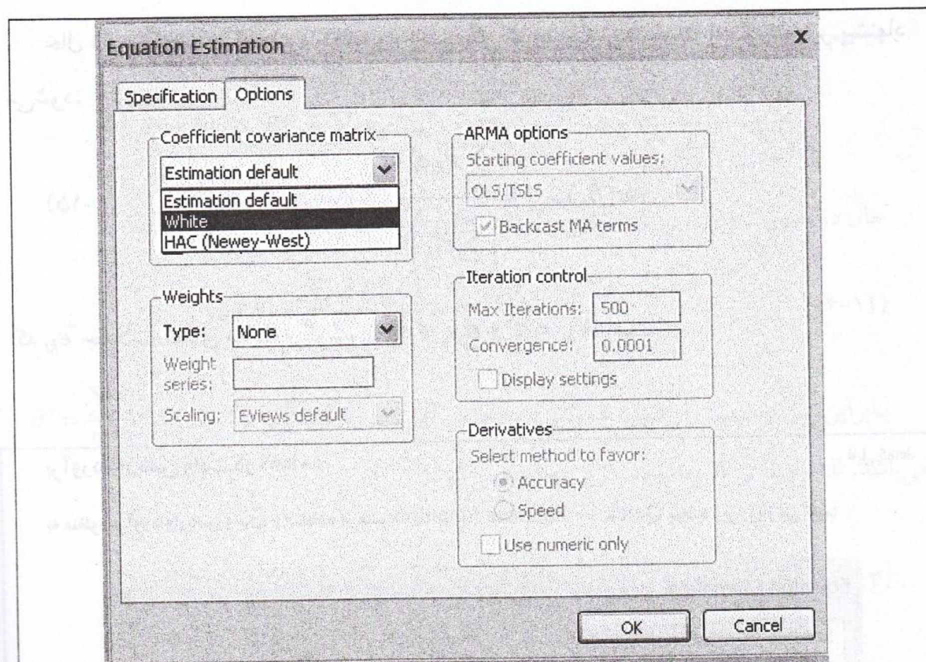
Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1952Q1 1996Q4

OK | Cancel

معادله موردنظر را در قسمت Equation Specification وارد می‌کنیم (به‌صورت Y C X2 X3). سپس با انتخاب گزینه options پنجرهٔ زیر را باز می‌کنیم:



در قسمت Coefficient Covariance matrix گزینه white را انتخاب می‌کنیم و معادله را تخمین می‌زنیم. این در حالی است که با انتخاب گزینه Estimation default واریانس‌های معمولی را حساب می‌کند. توجه شود که هرچند σ^2 هیچ تفاوتی نمی‌کند، ولی با انتخاب گزینه white، واریانس و انحراف معیار ضرایب متفاوت خواهد بود.

Equation: UNTITLED Workfile: DATA5::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y2
Method: Least Squares
Date: 12/26/13 Time: 13:49
Sample (adjusted): 1355 1386
Included observations: 32 after adjustments
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.348042	0.084660	98.60705	0.0000
X1	0.032141	0.007731	4.157231	0.0003
X2	0.027725	0.000237	117.2260	0.0000

R-squared	0.997950	Mean dependent var	9.614329
Adjusted R-squared	0.997808	S.D. dependent var	0.266791
S.E. of regression	0.012490	Akaike info criterion	-5.838762
Sum squared resid	0.004524	Schwarz criterion	-5.701349
Log likelihood	96.42019	Hannan-Quinn criter.	-5.793214
F-statistic	7057.919	Durbin-Watson stat	0.518669
Prob(F-statistic)	0.000000		

٤-٣-٤ آزمون‌های تشخیص واریانس ناهمسانی

آزمون‌های آماری مختلفی برای تشخیص واریانس ناهمسانی وجود دارد که در اینجا به برخی از این آزمون‌ها و نحوه انجام آنها با Eviews می‌پردازیم.

الف) آزمون بارتلت

آزمون بارتلت (۱۹۷۳) وقتی قابل کاربرد است که مشاهدات تکراری باشند. به‌عنوان مثال برای برآورد رابطه بین مصرف یک کالا با درآمد خانوار، مشاهدات تکراری وجود دارد. زیرا براساس آمار هزینه و درآمد خانوارها، مقادیر مصرف و درآمد هم بر حسب گروه‌های درآمدی و هم بر حسب سال موجود است. لذا اگر مصرف کالای Y را تابعی از درآمد (X) بدانیم، آنگاه در سال اول $n_۱$ خانوار (یا گروه درآمدی) داریم که برای آنها می‌توان معادله $Y_i = \alpha + \beta X_i + u_i$ را برآورد نمود. حال اگر t سال را در نظر بگیریم آنگاه معادله $Y_{it} = \alpha + \beta X_{it} + u_{it}$ را داریم که باقیمانده‌های آن e_{it} می‌باشند. حال برای خانوار iام (یا گروه درآمدی iام) T مشاهده داریم. بنابراین اگر معتقد باشیم که در این مدل، مشکل واریانس ناهمسانی وجود دارد و مسئلهٔ واریانس ناهمسانی بدین صورت است که با افزایش درآمد (X) واریانس‌ها افزایش می‌یابند، می‌توان برای هر خانوار یا گروه درآمدی، واریانس را به‌طور جداگانه حساب نمود. بدین منظور e_{it}ها را برای هر خانوار یا گروه درآمدی به‌صورت زیر نوشته و واریانس مربوطه را حساب می‌کنیم:

$$\text{گروه درآمدی ۱}: e_{۱۱}, e_{۱۲}, \ldots, e_{۱T} \Rightarrow \hat{\sigma}_۱^۲ = \frac{\sum_{t=۱}^{T} e_{۱t}^۲}{T-۱}$$

$$\text{گروه درآمدی ۲}: e_{۲۱}, e_{۲۲}, \ldots, e_{۲T} \Rightarrow \hat{\sigma}_۲^۲ = \frac{\sum_{t=۱}^{T} e_{۲t}^۲}{T-۱} \qquad (۴-۱۶)$$

$$\vdots$$

$$\text{گروه درآمدی } n: e_{n۱}, e_{n۲}, \ldots, e_{nT} \Rightarrow \hat{\sigma}_n^۲ = \frac{\sum_{t=۱}^{T} e_{nt}^۲}{T-۱}$$

اگر واریانس‌ها برابر باشند، آنگاه بایستی که $\hat{\sigma}_i^۲$ها به‌هم نزدیک باشند، در غیر این صورت دارای تفاوت‌های اساسی بوده که بیانگر واریانس ناهمسانی است. بارتلت آماره‌ای را برای آزمون

فرضیه $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$ ارائه می‌کند که شکل کلی آن برای حالتی که n گروه داشته و برای هر گروه T_i مشاهد داشته باشیم به‌صورت زیر است:

$$B = \frac{1}{f}\left[(N-n)\log\hat{\sigma}^2 - \sum_{i=1}^{n}(T_i - 1)\log\hat{\sigma}_i^2\right] \qquad (۴-۱۷)$$

$$f = 1 + \frac{1}{3(n-1)}\left[\sum_{i=1}^{n}\frac{1}{T_i - 1} - \frac{1}{N-1}\right] \quad ; \quad N = \sum_{i=1}^{n} T_i$$

$\hat{\sigma}^2$ میانگین واریانس‌ها است که برابر است با:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(T_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{n}(T_i - 1)} = \frac{\sum_{i=1}^{n}(T_i - 1)\hat{\sigma}_i^2}{N-n} \qquad (۴-۱۸)$$

اگر T_iها یکسان باشند، $\hat{\sigma}^2$ عبارت است از:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(T_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{n}(T_i - 1)} = \frac{(T-1)\sum_{i=1}^{n}\hat{\sigma}_i^2}{nT - n} = \frac{\sum_{i=1}^{n}\hat{\sigma}_i^2}{n-1} \qquad (۴-۱۹)$$

بدیهی است که اگر T_iها یکسان باشند، آنگاه آماره B به‌صورت زیر نوشته می‌شود (توجه شود که در این حالت $N = \sum T_i = nT$ است):

$$B = \frac{T-1}{f}\left(n\log\hat{\sigma}^2 - \sum_{i=1}^{n}\log\hat{\sigma}_i^2\right) \qquad (۴-۲۰)$$

$$f = 1 + \frac{1}{3(n-1)}\left[\frac{n}{T-1} - \frac{1}{nT-1}\right] = 1 + \frac{n+1}{3(T-1)}$$

B را می‌توان به‌صورت $\frac{T-1}{f}\log\frac{(\hat{\sigma}^2)^n}{\hat{\sigma}_1^2\,\hat{\sigma}_2^2\ldots\hat{\sigma}_n^2}$ نوشت. بدیهی است که اگر همهٔ واریانس‌ها برابر باشند آنگاه مخرج کسر نیز برابر با $(\hat{\sigma}^2)^n$ می‌شود که در این صورت، آماره B برابر صفر خواهد شد. در غیر این صورت، آماره B بزرگ شده و در ناحیه بحرانی قرار می‌گیرد که بیانگر رد H_0 (رد واریانس همسانی) می‌شود. آماره B توزیع χ^2 با درجه آزادی $n-1$ دارد.

ب) آزمون گلدفلد- کوانت[1]

در این آزمون فرض بر این است که مشاهدات را می‌توان به دو گروه تقسیم نمود و سپس آزمون یکسان بودن واریانس بین این دو گروه را انجام داد. مراحل این آزمون را به‌صورت زیر است:

۱- فرض کنید در رگرسیون $Y_t = \alpha + \beta X_t + u_t$ رابطهٔ $\sigma_t^2 = \sigma^2 X_t^2$ برقرار است. این بدان معنا است که عامل واریانس ناهمسانی، متغیر X_t است.

۲- مشاهدات را براساس مقادیر X_t از کوچک به بزرگ مرتب کرده و سپس آنها را به سه گروه تقسیم می‌کنیم:

- گروه ابتدایی با حجم n_1 که آن را گروه۱ می‌نامیم.
- گروه میانی با حجم m.
- گروه آخر با حجم n_2 که آن را گروه۲ می‌نامیم.

بدیهی است که $n_1 + m + n_2 = n$ می‌باشد.

۳- معادله رگرسیون را برای گروه اول و دوم برآورد کرده و RSSها را حساب می‌کنیم. مجموع مجذور خطاها را برای گروه اول و دوم به‌ترتیب با RSS_1 و RSS_2 نشان می‌دهیم.

۴- فرضیه $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ را آزمون می‌کنیم. این فرضیه بدان معنا است که واریانس گروه۱ و ۲ یکسان هستند. برای انجام این آزمون ابتدا $\hat{\sigma}_1^2$ و $\hat{\sigma}_2^2$ را حساب می‌کنیم که تحت فرضیه H_0، هر یک از آنها تخمین‌زنندهٔ σ^2 هستند. برای هر یک از این واریانس‌ها، روابط زیر برقرار است[2]:

$$\frac{\hat{\sigma}_1^2}{\sigma^2} = \frac{\chi^2_{n_1-K}}{n_1 - K} \quad , \quad \frac{\hat{\sigma}_2^2}{\sigma^2} = \frac{\chi^2_{n_2-K}}{n_2 - K} \qquad (۴-۲۱)$$

حال آماره F را برای مقایسه واریانس‌ها، تحت فرضیه H_0 تشکیل می‌دهیم:

$$F_{n_1-K, n_2-K} = \frac{\chi^2_{n_1-K} / n_1 - K}{\chi^2_{n_2-K} / n_2 - K} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{RSS_1 / n_1 - K}{RSS_2 / n_2 - K} = \frac{n_2 - K}{n_1 - K} \frac{RSS_1}{RSS_2} \qquad (۴-۲۲)$$

---

[1] Goldfeld–Quandt Test

[2] فصل اول بخش ۲-۷-۱ و ۴-۷-۱ را ببینید.

اگر مقدار F بزرگ باشد، در ناحیه بحرانی قرار می‌گیرد و فرضیه H_0 را رد می‌کند.

ج) آزمون گلجسر[1]

آزمون گلجسر قدر مطلق خطاها را روی متغیری که انتظار می‌رود واریانس ناهمسانی را توضیح می‌دهد (مانند Z) برازش می‌کند. شکل کلی چنین تابعی به‌صورت $|e_t| = \alpha_0 + \alpha_1 Z_t^{\gamma} + v_t$ است. از آنجا که این یک تابع غیرخطی است، لذا به ازای مقادیر مختلف γ برآورد می‌شود. معمولاً مقایری که برای γ در نظر گرفته می‌شود شامل ۱-، ۰/۵-، ۰/۵ و ۱ می‌باشد. به هر حال، معنادار بودن α_1 بیانگر رد فرضیه واریانس همسانی است.

د) آزمون همبستگی رتبه‌ای اسپیرمن[2]

اگر هدف ما بررسی رابطه بین واریانس u_t و X_t باشد می‌توان از ضریب همبستگی رتبه‌ای استفاده نمود. بدین منظور قدرمطلق e_tها را حساب کرده و سپس آنها و همچنین X_iها را رتبه‌بندی می‌کنیم. آنگاه تفاوت رتبه‌ها را حساب می‌کنیم:

$$X_t \text{ و } |e_t| \text{ تفاوت رتبه } = d_t = R_{|e_t|} - R_{X_t} \qquad (۴-۲۳)$$

$R_{|e_t|}$ و R_{X_t} به ترتیب رتبه‌های $|e_t|$ و X_t را نشان می‌دهند. ضریب همبستگی اسپیرمن را به‌صورت زیر محاسبه می‌شود:

$$r_s = 1 - 6\frac{\sum_{t=1}^{n} d_t^2}{n(n-1)} \qquad (۴-۲۴)$$

حال فرضیه عدم همبستگی بین X_t و e_t، یعنی فرضیه $H_0 : r_s = 0$ را آزمون می‌کنیم. فرضیه H_0 بیانگر واریانس همسانی است. معنی‌دار بودن ضریب همبستگی را می‌توان با آماره t آزمون نمود[1]:

---

[1]Golejser test
[2]Spearman's rank correlation test

$$t_{n-۲} = \frac{r_s}{\sqrt{\frac{۱ - r_s^۲}{n - ۲}}} \qquad (۴-۲۵)$$

اگر فرضیه $H_۰$ درست باشد، آنگاه $r_s = ۰$ بوده و آماره t صفر می‌شود. در غیر این صورت آماره t بزرگ بوده و در ناحیه بحرانی قرار می‌گیرد که فرضیه $H_۰$ را رد می‌کند. رد $H_۰$ بیانگر آن است که واریانس ناهمسانی وجود دارد.

هـ) آزمون بروش- پاگان-گادفری[۲]

آزمون بروش-پاگان برداری از متغیرها را که ممکن است موجب واریانس ناهمسانی شده باشند، در نظر می‌گیرد. شکل کلی این تابع به‌صورت زیر است:

$$\sigma_t^۲ = E(u_t^۲) = h(\alpha_۰ + \alpha_۱ Z_{۱t} + \cdots + \alpha_۲ Z_{rt}) \qquad (۴-۲۶)$$

Z_{it}ها می‌توانند دقیقاً با متغیرهای توضیحی (یعنی X_{it}ها) یکسان باشند یا متفاوت باشند. واریانس ناهمسانی توسط فرضیه $H_۰ : \alpha_۱ = \cdots = \alpha_r = ۰$ توصیف می‌شود. آزمون بروش-پاگان را می‌توان به‌صورت مراحل زیر انجام داد:

۱- ابتدا معادله $Y_t = \beta_۱ + \beta_۲ X_{۲t} + \cdots + \beta_K X_{Kt} + u_t$ را با روش OLS برآورد کرده و باقیمانده‌ها (e_t) را حساب می‌کنیم.

۲- واریانس معادله رگرسیون را به‌صورت $\hat{\sigma}^۲ = \frac{\sum e_t^۲}{n}$ حساب می‌کنیم.

۳- رگرسیون $\frac{e_t^۲}{\hat{\sigma}^۲} = \alpha_۰ + \alpha_۱ Z_{۱t} + \alpha_۲ Z_{۲t} + \cdots + \alpha_r Z_{rt} + \varepsilon_t$ را با روش OLS برآورد کرده و مجموع تغییرات توضیح داده شده (ESS_a) را حساب می‌کنیم.

۴- تحت فرضیه $H_۰$ رابطه زیر برقرار است:

$$\frac{۱}{۲} ESS_a \underset{asy}{\sim} \chi_{r-۱}^۲ \qquad (۴-۲۷)$$

---

[1] فصل دوم بخش ۱۰-۲ را ببینید.

[2] Breusch–Pagan test

بنابراین، اگر $\frac{ESS_a}{۲}$ بزرگتر از $\chi^{۲}_{\%.۰۵,r-۱}$ باشد، فرضیه $H_{\circ}$ (واریانس همسانی) رد می‌شود. زیرا در این حالت، متغیرهای Z_{it} توانسته‌اند $\sigma_t^{۲}$ یا $e_t^{۲}$ را توضیح دهند.

۵- از طرف دیگر می‌توان نشان داد که اگر $R_a^{۲}$ (به‌دست آمده از بند۳) را در تعداد مشاهدات ضرب کنیم، آنگاه $nR_a^{۲}$ توزیع $\chi^{۲}_{r-۱}$ دارد[1] که مشابه نتیجهٔ حاصل از بند ۴ است.

و) آزمون هاروی[2]

آزمون هاروی نیز مشابه آزمون‌های گلجسر و بروش-پاگان-گادفری است. در آزمون هاروی به‌جای $|e_t|$ و $e_t^{۲}$ از لگاریتم مجذور خطاها ($\log e_t^{۲}$) استفاده می‌شود. در این روش $\log e_t^{۲}$ روی متغیرهای توضیحی و یا هر متغیری که موجب واریانس ناهمسانی شده است، برازش می‌شود.

آزمون‌های واریانس ناهمسانی در Eviews

برای انجام آزمون واریانس ناهمسانی، در پنجره نتایج مسیر زیا را انتخاب می‌کنیم:

View → Residual Diagnostics → Heteroskedasticity Test

Equation: UNTITLED Workfile: ALMON::Demo\
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Representations
Estimation Output
Actual,Fitted,Residual
ARMA Structure...
Gradients and Derivatives
Covariance Matrix
Coefficient Diagnostics
Residual Diagnostics
Stability Diagnostics
Label

Std. Error	t-Statistic	Prob.
6.832761	-22.21996	0.0000
0.009089	175.7716	0.0000

Correlogram - Q-statistics...
Correlogram Squared Residuals...
Histogram - Normality Test
Serial Correlation LM Test...
Heteroskedasticity Tests...

Log likelihood	-919.5698
F-statistic	17391.98
Prob(F-statistic)	0.000000

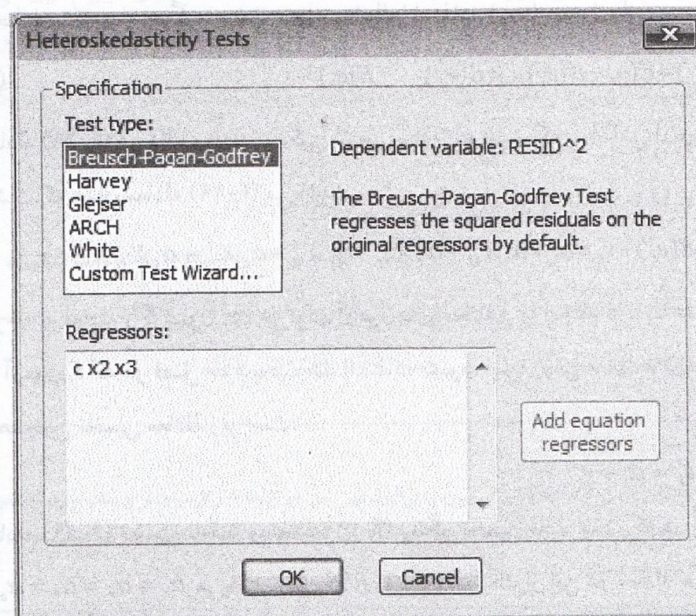
سپس در پنجره زیر نوع آزمون واریانس ناهمسانی را انتخاب می‌کنیم. آزمون‌های گادفری-بریش- پاگان، هاروی و گلجسر مشابه

[1] فصل هفتم بخش ۱۵-۷ را ببینید.

[2] Harvay

هم هستند و فقط متغیر وابسته آنها متفاوت است که به ترتیب به‌صورت e_t^2، $\log(e_t^2)$ و $|e_t^2|$ می‌باشند. در همه این آزمون‌ها در قسمت Regrossors هر متغیری که تصور می‌رود موجب واریانس ناهمسانی می‌شود وارد می‌کنیم.

اگر گزینه Arch را انتخاب کنیم نوع خاصی از واریانس ناهمسانی را آزمون می‌کند که جزئیات آن در فصل سیزدهم بحث شده است. گزینه White نیز آزمون وایت را انجام می‌دهد که در بخش بعدی بررسی شده است.



ز) آزمون وایت[1]

آزمون وایت حالت عمومی‌تری از آزمون بروش-پاگان-گادفری است. در آزمون وایت، فرضیه واریانس همسانی به‌صورت $H_0 : \sigma_t^2 = \sigma^2$ معرفی می‌شود. مراحل انجام این آزمون به‌صورت زیر است:

۱- معادله رگرسیون زیر را در نظر بگیرید:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (۴-۲۸)$$

برای آزمون « واریانس همسانی»، معادله (۴-۲۸) را برآورد کرده و باقیمانده‌ها (e_t) را حساب می‌کنیم.

[1] White

۲- معادله رگرسیون زیر را برآورد می‌کنیم:

$$e_t^2 = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{2t}^2 + \alpha_5 X_{3t}^2 + \alpha_6 X_{2t} X_{3t} + v_t \qquad (۴-۲۹)$$

v_t جملهٔ خطا است. در این رگرسیون از مجذور خطاها استفاده شده است، زیرا:

$$\text{var}(e_t) = E(e_2 - E(e_t))^2 = E(e_t^2) \quad ; \quad E(e_t) = 0 \qquad (۴-۳۰)$$

بنابراین، معادله (۲۹-۴) در واقع معادله میانگین شرطی برای e_t^2 است که بر اساس آن، رابطه زیر را می‌توان نوشت که همان معادله (۲۹-۴) می‌باشد:

$$e_t^2 = E(e_t^2) + v_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{2t}^2 + \alpha_5 X_{3t}^2 + \alpha_6 X_{2t} X_{3t} + v_t \qquad (۴-۳۱)$$

۳- با توجه به معادله رگرسیون e_t^2، می‌توان آزمون‌های مختلفی را انجام داد. به‌عنوان نمونه، می‌توان از آزمون F که در فصل سوم بحث شد استفاده نمود. این روش مستلزم تخمین معادله (۲۹-۴) و همچنین تخمین معادلهٔ زیر است:

$$e_t^2 = \alpha_1 + v_t \qquad (۴-۳۲)$$

در واقع معادله (۲۹-۴) تخمین نامقید و معادله(۳۲-۴) بیانگر تخمین مقید است که در آن ضرایب $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$ می‌باشد. حال RSS را برای هر یک از این دو معادله حساب کرده و F را برای آزمون آن، تشکیل می‌دهیم:

$$F = \frac{n-K}{m} \frac{RSS_R - RSS_{UR}}{RSS_{UR}} \qquad (۴-۳۳)$$

RSS_R و RSS_{UR} به‌ترتیب RSS مقید (معادله ۳۲-۴) و نامقید (معادله ۲۹-۴) می‌باشند. K تعداد ضرایب و m تعداد محدودیت‌ها است که در اینجا $K=6$ و $m=5$ می‌باشد. اگر اختلاف این دو رگرسیون زیاد باشد در این صورت مقدار F بزرگتر شده و بدین معنی است که فرضیه H_0 (یعنی ثابت بودن واریانس) رد می‌شود. بدیهی است که اگر $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$ باشد، بدین معنی است که واریانس u ثابت است.

روش دیگر برای آزمون واریانس همسانی معروف به آزمون ضریب لاگرانژ (LM)[1] است که نیازی به برآورد معادله (۳۲-۴) یعنی معادله رگرسیون مقید ندارد. در این روش از مقدار R^2 در

۱- Lagrange Multiplier. جزئیات این روش در فصل هفتم بخش ۱۵-۷ بحث شده است.

معادله (۲۹-۴) استفاده می‌شود. اگر یک یا چند ضریب در معادله (۲۹-۴) از نظر آماری معنی‌دار باشند، مقدار R^2 نسبتاً بزرگ خواهد بود. و در غیر این صورت مقدار R^2 کوچک خواهد بود. لذا آزمون LM بر اساس حاصل‌ضرب R^2 در تعداد مشاهدات است که دارای توزیع χ^2_m می‌باشد:

$$nR^2 \sim \chi^2_m \qquad (۴-۳۴)$$

m تعداد ضرایب تخمینی در معادله (۲۹-۴) (به‌استثنای جمله ثابت) است و از طرف دیگر برابر با تعداد محدودیت‌ها در آزمون F می‌باشد.

در آزمون LM، فرضیه مورد نظر به‌صورت زیر بیان می‌شود:

$$H_0: \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0 \qquad (۴-۳۵)$$

در آزمون LM اگر مقدار $\chi^2 = nR^2$ بزرگتر از مقدار بحرانی باشد، سپس فرضیه H_0 رد می‌شود و بدین معنی است که واریانس جمله خطا، ثابت نمی‌باشد.

آزمون وایت در Eviews — **فایل** data2

برای انجام آزمون وایت، ابتدا معادله رگرسیون را برآورد می‌کنیم. در اینجا از رگرسیون نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) استفاده می‌کنیم که روی رشد نیروی کار (GL) و رشد سرمایه (GK) و سهم مخارج مصرفی دولت از تولید ناخالص داخلی (SG) برازش شده است.

Equation: UNTITLED   Workfile: GOVERNMENT AND G...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GYNO
Method: Least Squares
Date: 01/23/11   Time: 18:54
Sample (adjusted): 1346 1379
Included observations: 34 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.060829	0.182337	5.817965	0.0000
GK	0.386531	0.087822	4.401316	0.0001
SG	0.039558	0.055057	0.718497	0.4778

R-squared	0.703292	Mean dependent var	5.645234
Adjusted R-squared	0.684149	S.D. dependent var	7.602924
S.E. of regression	4.272889	Akaike info criterion	5.826555
Sum squared resid	565.9849	Schwarz criterion	5.961233
Log likelihood	-96.05143	Hannan-Quinn criter.	5.872484
Durbin-Watson stat	1.591813		

در پنجره فوق، از منوی View مسیر زیر دنبال می‌کنیم:

View → Residual Tests → White Heteroskedasticity (no cross term)

توجه شود که no cross term بدین معنی است که حاصل‌ضرب متقاطع بین متغیرهای توضیحی را در نظر نمی‌گیرد (جملاتی مانند $X_{2t}X_{3t}$)، در حالی که اگر گزینه فوق را همراه با cross term انتخاب کنیم، حاصل‌ضرب‌های متقاطع را نیز در نظر می‌گیرد. نتایج آزمون وایت به‌صورت جدول زیر نشان داده می‌شود:

Equation: UNTITLED Workfile: GOVERNMENT AND GR...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: White

F-statistic	1.695178	Prob. F(9,24)	0.1452
Obs*R-squared	13.21369	Prob. Chi-Square(9)	0.1532
Scaled explained SS	11.62044	Prob. Chi-Square(9)	0.2356

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 01/23/11 Time: 19:41
Sample: 1346 1379
Included observations: 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-155.1392	113.2886	-1.369415	0.1835
GL	11.33672	6.814221	1.663685	0.1092
GL^2	-0.432020	0.247380	-1.746387	0.0935
GL*GK	0.110189	0.119207	0.924348	0.3645
GL*SG	-0.431259	0.333643	-1.292578	0.2085
GK	-1.736099	4.263417	-0.407208	0.6875
GK^2	0.109187	0.064568	1.691051	0.1038
GK*SG	0.052044	0.242705	0.214432	0.8320
SG	16.12752	12.74878	1.265025	0.2180
SG^2	-0.378120	0.353650	-1.069194	0.2956

R-squared	0.388638	Mean dependent var	16.52234
Adjusted R-squared	0.159377	S.D. dependent var	25.20729
S.E. of regression	23.11143	Akaike info criterion	9.358460
Sum squared resid	12819.32	Schwarz criterion	9.807390
Log likelihood	-149.0938	Hannan-Quinn criter.	9.511558
F-statistic	1.695178	Durbin-Watson stat	2.462716
Prob(F-statistic)	0.145172		

مقدار F و همچنین $\chi^2 = nR^2$ و احتمال‌های مربوطه نشان می‌دهد که این مدل، واریانس ناهمسانی ندارد. در واقع بر اساس F در سطح ۵ درصد، وجود واریانس ناهمسانی رد می‌شود. توجه شود که مقدار احتمال برای F برابر با ۱٤/٥۲ درصد به‌دست آمده است) در اینجا چون nR^2 نیز محاسبه شده است در قسمت بالای جدول با Obs*R-squared نشان داده شده است. لذا آزمون LM و وایت همراه با هم صورت گرفته است.

۵-۳-٤ تخمین ضرایب با وجود واریانس ناهمسانی

روش حداقل مربعات وزنی (WLS) یا روش حداقل مربعات تعمیم‌یافته (GLS)

برای توصیف روش GLS یا WLS، ابتدا حالت ساده‌ای را در نظر بگیرید که مدل K متغیره دارای واریانس ناهمسانی است. فرض کنید که واریانس u_t برابر با $\sigma_t^۲$ است که ناشی از X_{jt} می‌باشد.

$$Y_t = \beta_۱ + \beta_۲ X_{۲t} + \cdots + \beta_j X_{jt} + \cdots + \beta_K X_{Kt} + u_t \qquad (۴-۳۶)$$

$$\text{var}(u_t) = E(u_t^۲) = \sigma_t^۲ = \sigma^۲ X_{jt}^۲$$

برای حل مشکل واریانس ناهمسانی، طرفین معادله رگرسیون را بر X_{jt} تقسیم می‌کنیم (اگر طرفین را بر σ_t که برابر با σX_{jt} است تقسیم کنیم، همین نتیجه به‌دست می‌آید):

$$\frac{Y_t}{X_{jt}} = \beta_۱ \frac{۱}{X_{jt}} + \beta_۲ \frac{X_{۲t}}{X_{jt}} + \cdots + \beta_j + \cdots + \beta_K \frac{X_{Kt}}{X_{jt}} + \frac{u_t}{X_{jt}}$$

متغیرهای فوق را که بر X_{jt} تقسیم شده‌اند با علامت ستاره و همچنین $\frac{۱}{X_{jt}}$ را با $X_{۱t}^*$ نشان می‌دهیم:

$$Y_t^* = \beta_j + \beta_۱ X_{۱t}^* + \beta_۲ X_{۲t}^* + \cdots + \beta_K X_{Kt}^* + u_t^* \qquad (۴-۳۷)$$

می‌توان نشان داد که مدل فوق دارای واریانس همسانی است:

$$\text{var}(u_t^*) = E(u_t^{*۲}) = E(\frac{u_t^۲}{X_{jt}^۲}) = \frac{E(u_t^۲)}{X_{jt}^۲} = \frac{\sigma_t^۲}{X_{jt}^۲} = \frac{\sigma^۲ X_{jt}^۲}{X_{jt}^۲} = \sigma^۲$$

بنابراین می‌توان مدل فوق را با روش OLS تخمین زد، زیرا واریانس ناهمسانی ندارد.

در حالت کلی، شکل ماتریس معادله K متغیره به‌صورت زیر است.

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \qquad (۴-۳۸)$$

ماتریس واریانس $\mathbf{u}$ عبارت است از:

$$\mathbf{\Omega} = \text{var}(\mathbf{u}) = E(\mathbf{uu'}) = E \begin{bmatrix} u_۱ \\ u_۲ \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} u_۱ & u_۲ & \cdots & u_n \end{bmatrix}$$

$$= E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \vdots & \vdots & & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix} \qquad (۴-۳۹)$$

اگر u_tها خودهمبستگی نداشته باشند آنگاه $\sigma_{ij} = 0$ خواهد بود و لذا $\mathbf{\Omega}$ یک ماتریس قطری خواهد بود.

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \qquad (۴-۴۰)$$

اگر واریانس همسانی برقرار باشد، آنگاه $\mathbf{\Omega}$ برابر است با:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \sigma^2\mathbf{I}_n \qquad (۴-۴۱)$$

از طرف دیگر، تخمین‌زنندهٔ OLS برای $\boldsymbol{\beta}$ و σ^2 به‌صورت زیر است[1]:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (۴-۴۲)$$

$$\hat{\sigma}^2 = \frac{e'e}{n-K} = \frac{\sum e_t^2}{n-K} \quad , \quad \hat{\mathbf{\Omega}} = \hat{\sigma}^2\mathbf{I}_n \qquad (۴-۴۳)$$

اگر از (۴-۳۸) به‌جای $\mathbf{y}$ در (۴-۴۲) جایگذاری کنیم به $\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ می‌رسیم. حال واریانس $\hat{\boldsymbol{\beta}}_{OLS}$ برابر است با:

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}) &= E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \qquad (۴-۴۴)$$

بنابراین اگر شرط واریانس همسانی برقرار باشد، در این صورت $\mathbf{\Omega} = \sigma^2\mathbf{I}_n$ است و لذا توزیع تخمین‌زنندهٔ $\hat{\boldsymbol{\beta}}_{OLS}$ عبارت است از:

---

[1] فصل سوم بخش ۱۰-۳ را ببینید.

$$\hat{\boldsymbol{\beta}}_{OLS} \sim \mathrm{N}\left(\boldsymbol{\beta}\,,\,\sigma^{2}(\mathbf{X'X})^{-1}\right) \qquad (۴-۴۵)$$

اما اگر واریانس همسانی برقرار نباشد، در این صورت توزیع $\hat{\boldsymbol{\beta}}$ با روش OLS عبارت است از:

$$\hat{\boldsymbol{\beta}}_{OLS} \sim \mathrm{N}\left(\boldsymbol{\beta}, (\mathbf{X'X})^{-1}(\mathbf{X'\Omega X})(\mathbf{X'X})^{-1}\right) \qquad (۴-۴۶)$$

مقایسه نتایج (۴۵-۴) و (۴۶-۴) نشان می‌دهد که اگر مدل رگرسیون دارای واریانس ناهمسانی باشد و آن را نادیده بگیریم، در این صورت تخمین واریانس‌ها به‌درستی انجام نخواهد شد. برای حل این مشکل، ابتدا بایستی مدل را به‌گونه‌ای تبدیل کنیم که واریانس ناهمسانی آن برطرف شود و سپس $\hat{\boldsymbol{\beta}}$ را تخمین بزنیم. این روش معروف به حداقل مربعات تعمیم‌یافته (GLS)[1] یا حداقل مربعات وزنی (WLS)[2] است.

به منظور استفاده از روش حداقل مربعات تعمیم‌یافته، مشابه معادله (۳۷-۴) عمل می‌کنیم که طرفین آن را بر عاملی که موجب واریانس ناهمسانی شده است، تقسیم می‌کنیم. در اینجا بحث مذکور را در حالت عمومی به‌کار می‌بریم. بدین منظور توجه کنیم که اگر ماتریس $\mathbf{\Omega}$ مثبت معین باشد می‌توان آن را به‌صورت $\mathbf{\Omega}^{-1} = \mathbf{P'P}$ نوشت که در اینجا $\mathbf{P} = \mathbf{\Omega}^{-\frac{1}{2}}$ است. از طرف دیگر، چون $\mathbf{\Omega}$ متقارن است لذا $\mathbf{P} = \mathbf{P'}$ است. حال طرفین معادله را در $\mathbf{P}$ ضرب می‌کنیم (در واقع آن را بر عامل واریانس ناهمسانی تقسیم کرده‌ایم، زیرا $\mathbf{P} = \mathbf{\Omega}^{-\frac{1}{2}}$ است).

$$\mathbf{Py} = \mathbf{PX\beta} + \mathbf{Pu} \qquad (۴-۴۷)$$

مدل فوق را به‌صورت زیر می‌نویسیم:

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \mathbf{u}_* \;\; ; \;\; \mathbf{y}_* = \mathbf{Py} \;\; , \;\; \mathbf{X}_* = \mathbf{PX} \;\; , \;\; \mathbf{u}_* = \mathbf{Pu} \qquad (۴-۴۸)$$

$$\mathbf{y}_* = \mathbf{Py} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \frac{Y_1}{\sigma_1} \\ \frac{Y_2}{\sigma_n} \\ \vdots \\ \frac{Y_n}{\sigma_n} \end{bmatrix}$$

[1] Generalized Least Squares
[2] Weighted Least Squares

$$\mathbf{X}_* = \mathbf{PX} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix} \begin{bmatrix} 1 & X_{21} & \cdots & X_{K1} \\ 1 & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & & \vdots \\ 0 & X_{2n} & \cdots & X_{Kn} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1} & \frac{X_{21}}{\sigma_1} & \cdots & \frac{X_{K1}}{\sigma_1} \\ \frac{1}{\sigma_2} & \frac{X_{22}}{\sigma_2} & \cdots & \frac{X_{K2}}{\sigma_2} \\ \vdots & \vdots & & \vdots \\ \frac{1}{\sigma_n} & \frac{X_{2n}}{\sigma_n} & \cdots & \frac{X_{Kn}}{\sigma_n} \end{bmatrix}$$

$$\mathbf{u}_* = \mathbf{Pu} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} \frac{u_1}{\sigma_1} \\ \frac{u_2}{\sigma_1} \\ \vdots \\ \frac{u_n}{\sigma_n} \end{bmatrix}$$

تخمین $\boldsymbol{\beta}$ برای مدل (۴-۴۸) معروف به تخمین‌زنندهٔ GLS است که به‌صورت زیر به‌دست می‌آید:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}(\mathbf{X}_*'\mathbf{y}_*) = [(\mathbf{PX})'(\mathbf{PX})]^{-1}[(\mathbf{PX})'(\mathbf{Py})] = (\mathbf{X}'\mathbf{P}'\mathbf{PX})^{-1}(\mathbf{X}'\mathbf{P}'\mathbf{Py}) = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}) \quad (۴-۴۹)$$

تخمین دو مرحله‌ای

تا اینجا دیدیم که تخمین‌زنندهٔ $\hat{\boldsymbol{\beta}}$ در صورت واریانس ناهمسانی برابر با (۴-۴۹) است که $\boldsymbol{\Omega}$ برابر است با:

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

استفاده از (۴-۴۹) برای محاسبه $\hat{\boldsymbol{\beta}}$ نیازمند مقادیر $\mathbf{X}$ ، $\mathbf{y}$ و $\boldsymbol{\Omega}$ است. به‌هرحال مقادیر $\mathbf{X}$ و $\mathbf{y}$ را داریم اما $\boldsymbol{\Omega}$ و عناصر آن یعنی σ_i^2 مجهول است. لذا بایستی تخمین آنها یعنی $\hat{\sigma}_i^2$ را داشته باشیم تا با استفاده از $\hat{\boldsymbol{\Omega}}$ بتوانیم تخمین $\hat{\boldsymbol{\beta}}$ را به‌دست آوریم. این روش معروف به روش حداقل مربعات تعمیم‌یافتهٔ قابل دسترس (FGLS)[1] است.

---

[1]Feasible Generalized Least Squares

برای محاسبه واریانس‌ها نیاز به باقیمانده‌ها (خطاها) داریم. تصور کنید که $\sigma_t^2 = \mathbf{z}_t'\boldsymbol{\alpha}$ است که $\boldsymbol{\alpha}$ بردار ضرایب و $\mathbf{z}_t$ بردار متغیرهایی است که موجب واریانس ناهمسانی شده‌اند ($\mathbf{z}_t$ ممکن است شامل $\mathbf{x}_t$ نیز باشد). در این صورت، ماتریس $\boldsymbol{\Omega}$ وابسته به ضرایب $\boldsymbol{\alpha}$ است که بایستی آنها را برآورد کنیم. از طرف دیگر اگر e_t جملات خطای مدل (۴-۳۶) یا (۴-۳۸) باشد، در این صورت، با استفاده از رابطه $E(e_t^2) = \sigma_t^2$ خواهیم داشت:

$$e_t^2 = \sigma_t^2 + v_t = \mathbf{z}_t'\boldsymbol{\alpha} + v_t \qquad (۴-۵۰)$$

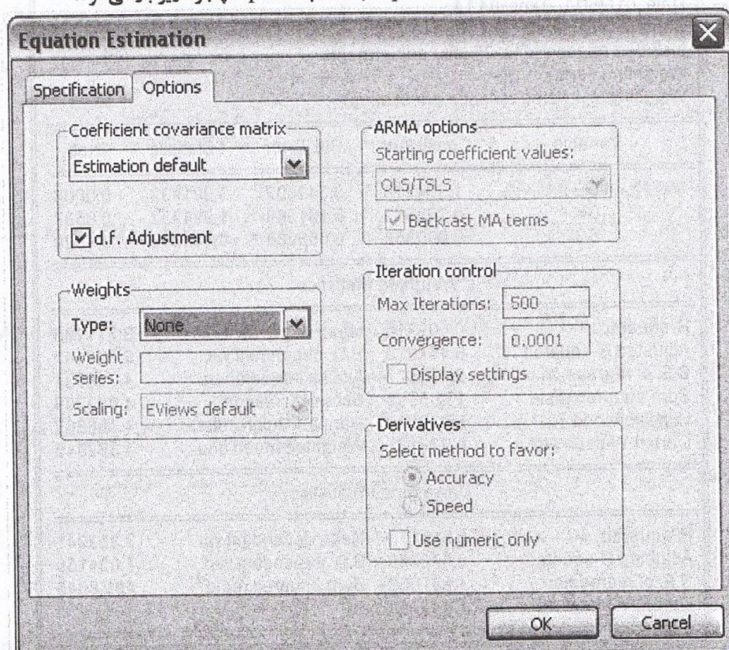
با برآورد مدل فوق، $\hat{\boldsymbol{\alpha}}$ به‌دست می‌آید و با استفاده از $\hat{\sigma}_t^2 = \mathbf{z}_t'\hat{\boldsymbol{\alpha}}$ را حساب کرده و ماتریس $\hat{\boldsymbol{\Omega}}$ را تشکیل می‌دهیم. لذا تخمین‌زنندهٔ FGLS عبارت است از:

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}) \qquad (۴-۵۱)$$

روش دیگر برای برآورد ضرایب، استفاده از روش حداکثر درستنمایی است که در فصل هفتم بحث شده است.

روش GLS در Eviews — **فایل** data2

با تخمین معادله موردنظر، در پنجرهٔ Equation Estimation با انتخاب Option پنجرهٔ زیر باز می‌شود:



در این پنجره در قسمت Type نوع وزنی را که به داده‌ها می‌خواهیم بدهیم، مشخص می‌کنیم. به‌عنوان مثال اگر بخـواهیم وزنها را متناسب با واریانس بدهیم بایستی گزینه Variance را انتخاب کنیم. همچنین اگر بخواهیم از معکوس واریانس بـه‌عنـوان وزن استفاده کنیم بایستی از گزینه Inverse Variance استفاده کنیم. مشابه این دو گزینه برای انحراف معیار نیز وجود دارد.

None
Inverse std. dev.
Inverse variance
Std. deviation
Variance

در مرحله بعد بایستی وزن مورد نظـر را در قـسمت Weight Series وارد کنـیم. بـه‌عنـوان مثـال اگـر گزینـه Inverse variance را انتخاب و 1/X را به‌عنوان وزن بدهیم، در این صورت مشابه آن است که طرفین معادله رگرسـیون را بـر X تقـسیم کرده‌ایم.

در قسمت Scaling سه گزینه وجود دارد. گزینه Average بدان معنا است که Eviews وزن‌ها را به‌گونه‌ای مقیاس‌بندی می‌کند که جمع آنها برابر با تعداد مشاهدات (T) گردد. گزینه Eviews default مقیاس وزن‌ها را طوری تعیین مـی‌کنـد کـه حاصل‌جمع جذر وزن‌ها برابر با تعداد مشاهدات (T) گردد.

فرض کنید در معادله نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) به ایـن نتیجـه رسیده‌ایـم کـه دارای واریـانس ناهمسانی است. همچنین فرض کنید نتیجه گرفته‌ایم که واریانس ناهمسانی ناشی از متغیر GL است. بدین منظور طرفین معادله را بر GL تقسیم کرده و آن را تخمین می‌زنیم. برای استفاده از روش WLS ، طبق آنچه که گفتـه شـد، در قـسمت Weight series عبارت 1/GL و در قسمت Type گزینه Inverse variance را انتخاب می‌کنیم. نتایج تخمین عبارت است از:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GYNO
Method: Least Squares
Date: 01/25/11 Time: 04:14
Sample (adjusted): 1350 1380
Included observations: 25 after adjustments
Weighting series: 1/GL
Weight type: Inverse variance (no scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.364219	0.342602	3.981932	0.0006
GK	0.202553	0.101569	1.994232	0.0587
SG	0.030358	0.068688	0.441963	0.6628

Weighted Statistics

R-squared	0.502315	Mean dependent var	3.711109
Adjusted R-squared	0.457070	S.D. dependent var	2.730447
S.E. of regression	2.454677	Akaike info criterion	4.746034
Sum squared resid	132.5596	Schwarz criterion	4.892299
Log likelihood	-56.32542	Hannan-Quinn criter.	4.786602
Durbin-Watson stat	0.928161	Weighted mean dep.	4.362945

Unweighted Statistics

R-squared	0.611113	Mean dependent var	7.353941
Adjusted R-squared	0.575760	S.D. dependent var	7.034155
S.E. of regression	4.581606	Sum squared resid	461.8045
Durbin-Watson stat	1.042760		

٤-٤ فرض سوم: عدم خودهمبستگی[1]

عدم خودهمبستگی بین جملات خطا یکی از فروض اساسی برای روش OLS است. همبستگی خطاها با یکدیگر را اصطلاحاً «خودهمبستگی» یا «همبستگی سریالی» می گویند. برای بررسی خودهمبتگی، از برآورد خطاها (یعنی e_t) استفاده می‌شود. به‌طور کلی خودهمبستگی بیانگر رابطه بین e_t با وقفه‌های آن است.

١-٤-٤ مفهوم وقفه[2]

مقدار وقفهٔ هر متغیری مانند X_t برابر با مقدار آن در زمان قبلی(X_{t-1}) می‌باشد. به‌عنوان مثال مقادیر X_t و وقفه آن را می‌توان در جدول زیر نشان داد:

زمان	X_t	X_{t-1}	$\Delta X_t = X_t - X_{t-1}$
۱۳۸۰	۱	—	—
۱۳۸۱	۲	۱	۲-۱=۱
۱۳۸۲	۴	۲	۴-۲=۲
۱۳۸۳	-۲	۴	-۲-۴=-۶
۱۳۸۴	-۳	-۲	-۳-(-۲)=-۱
۱۳۸۵	۰	-۳	۰-(-۳)=۳
۱۳۸۶	۵	۰	۵-۰=۵

به‌عنوان مثال در سال ۱۳۸۲، مقدار X در این سال را مقدار جاری و مقدار آن در سال قبل (سال ۱۳۸۱) را مقدار X با «یک وقفه» می گویند. تفاضل این دو مقدار (ΔX_t) بیانگر تفاضل مرتبه اول است که مقدار تغییرات X در سال جاری نسبت به سال قبل را نشان می‌دهد.

١-٤-٤ پیامدهای خودهمبستگی

اگر معادله $Y_t = \alpha + \beta X_t + u_t$ دارای خودهمبستگی به‌صورت $u_t = \rho u_{t-1} + \varepsilon_t$ باشد ولی آن را نادیده بگیریم، پیامدهای زیر را خواهد داشت:

۱- تخمین‌زننده‌های OLS همچنان بدون تورش هستند، زیرا این موضوع مربوط به فرض اول ($E(u_t) = 0$) است که خودهمبستگی موجب نقض آن نمی‌شود.

۲- تخمین‌زننده‌های OLS سازگارند، زیرا این موضوع مربوط به فرض غیرتصادفی بودن X_t یا عدم همبستگی بین u_t و X_t است که خودهمبستگی موجب نقض آن نمی‌شود. این بدان معنا است که با افزایش حجم نمونه، واریانس $\hat{\beta}$ به صفر گرایش دارد.

---

[1] autocorrelation
[2] lag

۳- تخمین‌زننده OLS کارا نخواهد بود و واریانس $\hat{\beta}$ برابر با $\frac{\sigma^2}{\sum x_t^2}$ نمی‌باشد، زیرا:

$$\operatorname{var}(\hat{\beta}) = E(\hat{\beta}-\beta)^2 = E\left(\sum_t w_t u_t\right)^2 \quad , \quad w_t = \frac{x_t}{\sum x_t^2}$$

$$= E(\sum_t w_t^2 u_t^2) + E(2\sum_t \sum_{\substack{s \\ t>s}} w_t w_s u_t u_s)$$

$$= \sum_t w_t^2 E(u_t^2) + 2\sum_t \sum_{\substack{s \\ t>s}} w_t w_s E(u_t u_s)$$

با توجه‌به $E(u_t u_s) = \rho^{t-s}\sigma^2$ و $E(u_t^2) = \sigma^2$، خواهیم داشت(ρ^{t-s} ضریب همبستگی u_t و u_s است)[1]:

$$\operatorname{var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_t^2} + 2\sum_t \sum_{\substack{s \\ t>s}} w_t w_s \rho^{t-s} \sigma^2 \qquad (۴-۵۲)$$

بنابراین، واریانس $\hat{\beta}$ برابر با $\frac{\sigma^2}{\sum x_t^2}$ نخواهد بود. لذا برآوردهای OLS منجر به تورش در تخمین واریانس $\hat{\beta}$ می‌شود که مقدار آن بستگی به حاصل‌جمع دوم در سمت راست رابطه فوق دارد. اگر u_tها خودهمبستگی مثبت (یعنی $0 < \rho < 1$) و X_tها نیز خودهمبستگی مثبت داشته باشند[2]، آنگاه حاصل جمع مذکور، مثبت بوده و لذا واریانس $\hat{\beta}$ در روش OLS کمتر از حد، برآورد می‌شود. این نیز به‌نوبه خود موجب بزرگ شدن tها خواهد شد.

۴-٤-۲ روش‌های نموداری جهت تشخیص خودهمبستگی

یکی از راه‌های ساده جهت تشخیص خودهمبستگی، استفاده از نمودار خطاها (e_t) است. بدین منظور برای بررسی رابطه بین e_t و e_{t-1}، می‌توان e_t را در مقابل e_{t-1} ترسیم نمود. با ترسیم این داده‌ها، حالت‌های مختلفی را می‌توان مشاهده نمود. به‌عنوان مثال اگر نموداری مانند (۳-۴) مشاهده کنیم، بیانگر وجود خودهمبستگی مثبت می‌باشد[3].

---

[1] برای اثبات این رابطه، می‌توان از تعاریف زیر استفاده نمود:

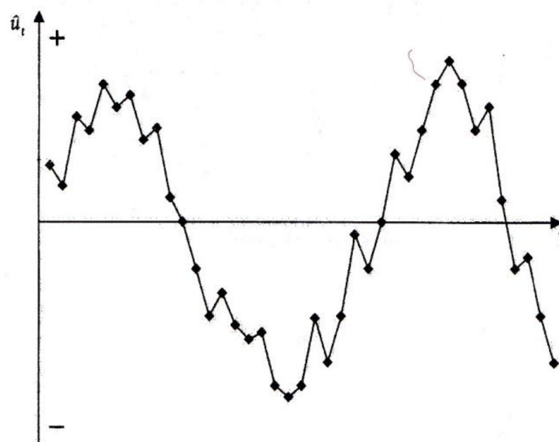
$$\rho^{t-s} = \frac{\operatorname{cov}(u_t, u_s)}{\sqrt{\operatorname{var}(u_t)}\sqrt{\operatorname{var}(u_s)}} = \frac{E(u_t u_s)}{\sigma^2}$$

زیرا $\operatorname{var}(u_t) = \operatorname{var}(u_s) = \sigma^2$ است. به‌عنوان مثال $\rho = \frac{E(u_t u_{t-1})}{\sigma^2}$ و $\rho^2 = \frac{E(u_t u_{t-2})}{\sigma^2}$ است.

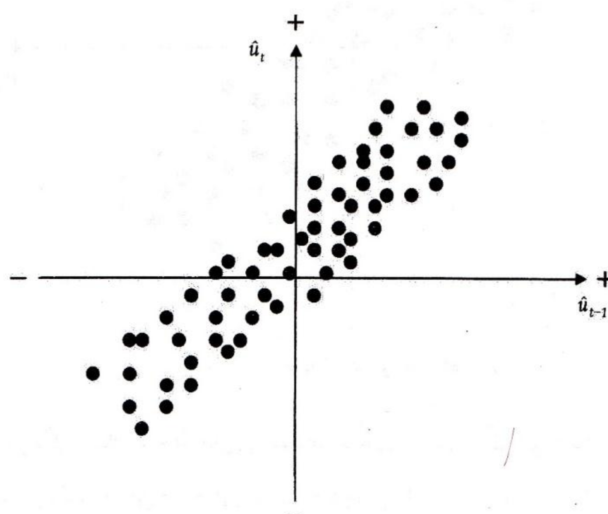
[2] این بدان معنا است که $\sum w_t w_s = \frac{\sum x_t x_s}{(\sum x_t^2)^2} > 0$ باشد، زیرا صورت کسر بیانگر $\operatorname{cov}(x_t, x_s)$ است.

[3] این نمودارها از Brooks (2008) می‌باشد.

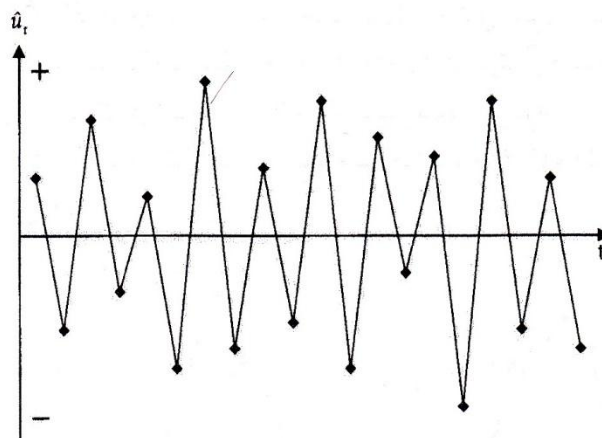
از طرف دیگر اگر مشاهدات مربوط به خطاها (e) را در مقابل زمان ترسیم کنیم، نمودار (۲-۴) به دست خواهد آمد. نمودارهای (۲-۴) و (۳-۴) دقیقاً یکسان هستند. در واقع «خودهمبستگی مثبت» که از نمودار (۳-۴) مشاهده می‌شود، می‌توان آن را در نمودار (۲-۴) به‌صورت تغییرات منظم e در طول زمان، مشاهده نمود. نمودار (۴-۴) و (۵-۴) خودهمبستگی منفی را نشان می‌دهد.



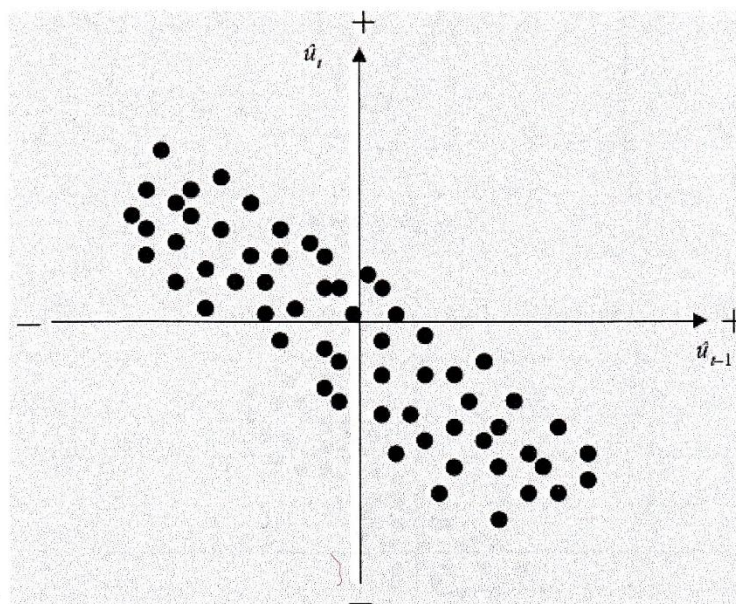
نمودار ۲-٤: خودهمبستگی مثبت



نمودار ۳-٤: خودهمبستگی مثبت



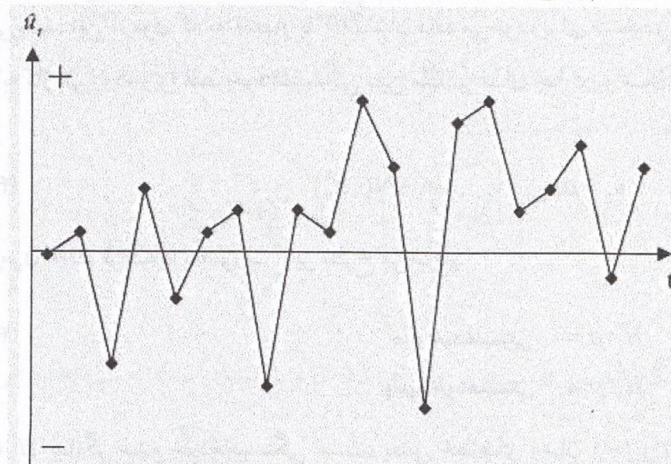
نمودار ٤-٤: خودهمبستگی منفی



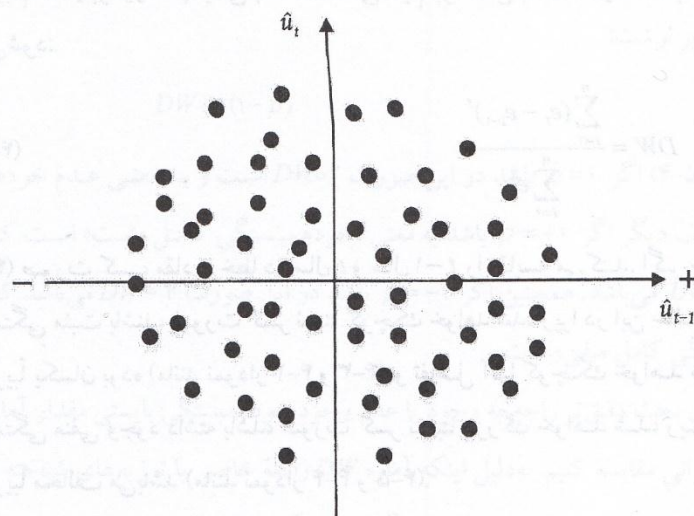
نمودار ٥-٤: خودهمبستگی منفی

بنابراین اگر جملات خطا دارای خودهمبستگی باشند بیانگر این است که در طول زمان به‌طور منظم تغییر می‌کنند و این در حالی است که بایستی تغییرات آنها کاملاً تصادفی باشد. مثالی از تصادفی بودن جملات خطا در نمودارهای (۶-۴) و (۷-۴) نشان داده شده است. نمودار (۶-۴)

نشان می‌دهد که تغییرات جملات خطا در طول زمان از هیچ قاعده خاصی تبعیت نمی کند و کاملاً تصادفی است. نمودار (۷-۴) نیز نشان می‌دهد که بین جملات خطا در دوره فعلی و دوره قبلی هیچ ارتباط خاصی وجود ندارد و اگر بین u_t و u_{t-1} ضریب همبستگی را حساب کنیم و یا رگرسیون بین آنها را برآورد نماییم، هیچ رابطه معناداری به دست نخواهد آمد.



نمودار ۶-۴: عدم خودهمبستگی



نمودار ۷-۴: عدم خودهمبستگی

۳-٤-٤ آزمون دوربین- واتسون

قدم اول در شناسایی خودهمبستگی این است که نمودار آنها را به شرحی که گفته شد، ترسیم کنیم. اما در عمل، روش‌های نموداری و تحلیل آنها مشکل است و لـذا برخـی روش‌هـای آمـاری جهت آزمون خودهمبستگی ارائه شده است. یکی از آزمون‌های متداول، آزمون دوربین- واتسون (۱۹۵۱) می‌باشد. این آزمون که به‌اختصار با DW نشان داده می‌شود، برای تشخیص خودهمبستگیِ مرتبه اول به‌کار می‌رود، زیرا فقط خودهمبستگی بـین مقـادیر سـال جـاری و سـال قبـل را در نظـر می‌گیرد:

$$u_t = \rho u_{t-1} + v_t \quad , \quad v_t \sim N(0, \sigma_v^2) \qquad (۴-۵۳)$$

در آزمون DW، فرضیه‌ها به‌صورت زیر مطرح می‌شوند:

$$H_0 : \rho = 0 \quad \text{عدم خودهمبستگی} \qquad (۴-۵۴)$$
$$H_1 : \rho \neq 0 \quad \text{وجود خودهمبستگی}$$

فرضیه H_0 بیـانگر عـدم خودهمبـستگی اسـت، یعنـی خطاهـای زمـان $t-1$ و t مـستقل از هـم می‌باشند که در این صورت $u_t = v_t$ بوده و لذا u_t کاملاً تصادفی است. از آنجا که اطلاعاتی راجع به u_t نداریم لذا از برآورد آنها یعنی e_t استفاده می‌کنیم. بر اساس e_tها، آماره DW به‌صورت زیـر تعریف می‌شود:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2} \qquad (۴-۵۵)$$

در (۴-۵۵) صورت کسر، مقادیر خطا در سال t و سال $t-1$ را مقایسه می‌کنـد. اگـر خطاهـا دارای خودهمبستگیِ مثبت باشند، صورت کسر نسبتاً کوچک خواهد شد، زیرا در این حالت، علامت e_t و e_{t-1} تقریباً یکسان بوده (مانند نمودار ۴-۲ و ۴-۳) و تفاضل آنها کوچـک خواهـد شـد. امـا اگـر خودهمبستگی منفی وجود داشته باشد، صورت کسر نـسبتاً بـزرگ خواهـد شـد، زیـرا علامـت e_t و e_{t-1} تقریباً مخالف می‌باشد (مانند نمودار ۴-۴ و ۴-۵).

برای ساده نمودن DW، صورت کسر را بازنویسی می‌کنیم:

$$\sum_{t=2}^{n}(e_t - e_{t-1})^2 = \sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}$$

اما از آنجا که اختلاف چندانی بین $\sum e_t^2$ و $\sum e_{t-1}^2$ وجود ندارد، لذا تقریباً یکسان هستند:

$$\sum_{t=2}^{n} e_t^2 = e_2^2 + e_3^2 + \dots + e_n^2 \cong \sum_{t=2}^{n} e_{t-1}^2 = e_1^2 + e_2^2 + \dots + e_{n-1}^2$$

تفاوت این دو عبارت در این است که اولی e_1^2 و دومی e_n^2 را ندارد. به‌هرحال با یکسان گرفتن این دو عبارت، خواهیم داشت:

$$DW \cong \frac{2\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2} = 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}\right)$$

از طرف دیگر، تخمین ρ بر اساس معادله (۲۲-۴) و با استفاده از e_t به‌جای u_t، برابر است با:

$$\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$$

$\hat{\rho}$ ضریب همبستگی e_t و e_{t-1} را نشان می‌دهد که برآورد ρ است. حال می‌توان DW را به‌صورت زیر نوشت:

$$DW \cong 2(1 - \hat{\rho}) \qquad (۴-۵۶)$$

طبق (۵۶-۴) اگر $\hat{\rho} = 0$ باشد در این صورت $DW=2$ است و به معنی عدم خودهمبستگی است. از طرف دیگر اگر $\rho = +1$ باشد به معنی «خودهمبستگی کامل مثبت» است که در این حالت $DW = 0$ می‌باشد. همچنین اگر $\rho = -1$ باشد در این صورت $DW = 4$ می‌باشد که به معنی «خودهمبستگی کامل منفی» است.

اما برای بحث دقیق‌تر راجع به وجود یا عدم وجود خودهمبستگی بایستی مقدار آماره DW را با مقادیر بحرانی مقایسه کنیم. به‌دلیل اینکه آماره DW رابطهٔ خاصی با توزیع‌های شناخته شده مانند t، F و χ^2 ندارد، لذا برای DW جداول جداگانه‌ای تشکیل شده است. در این جدول K تعداد

ضرایب به استثنای عرض از مبدأ است و n تعداد مشاهدات می‌باشد. قضاوت راجع به وضعیت خودهمبستگی طبق نمودار زیر می‌باشد:

خودهمبستگی منفی (رد H_0)	نمی‌توان نتیجه‌گیری کرد	عدم خودهمبستگی (H_0 رد نمی‌شود)	نمی‌توان نتیجه‌گیری کرد	خودهمبستگی مثبت (رد H_0)

٤ ، $4-d_L$ ، $4-d_U$ ، ۲ ، d_U ، d_L ، ۰

بدین ترتیب اگر DW به اندازه کافی به ۲ نزدیک باشد، به معنی عدم خودهمبستگی است. اما اگر به صفر نزدیک باشد، خودهمبستگیِ مثبت و اگر به ۴ نزدیک باشد، خودهمبستگیِ منفی را نشان می‌دهد.

مثال ١-٤: فرض کنید که مدل زیر با $n=۸۰$ مشاهده برآورد شده است:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

فرض کنید که برای این مدل $DW = 1/25$ به دست آمده باشد. برای قضاوت راجع به خودهمبستگی، مقادیر زیر با توجه به اینکه تعداد ضرایب بااستثنای عرض از مبدأ برابر ۳ و $n=۸۰$ است، از جدول آماره دوربین-واتسون به دست آمده است:

$$d_L = 1/56 \quad , \quad d_U = 1/72$$

$$4 - d_U = 2/28 \quad , \quad 4 - d_L = 2/44$$

چون $DW = 1/25$ کمتر از $d_L = 1/56$ می‌باشد لذا این معادله دارای خودهمبستگی مثبت می‌باشد.

اما برای اینکه آزمون DW معتبر باشد، بایستی شرایط زیر برقرار باشد:

۱- معادله رگرسیون بایستی دارای جمله ثابت باشد.

۲- متغیرهای توضیحی بایستی غیرتصادفی باشند.

۳- متغیر وابسته تأخیری نباید در مدل وارد شده باشد.

در رابطه با مشکل سوم فرض کنید معادله رگرسیون به‌صورت زیر باشد:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \theta Y_{t-1} + u_t \qquad (۴-۵۷)$$

در این صورت برای بررسی خودهمبستگی از آماره دیگری به نام h دوربین استفاده می‌شود که عبارت است از:

$$h = \hat{\rho}\sqrt{\frac{n}{۱-n\,\text{var}(\hat{\theta})}} \quad , \quad \hat{\rho} = ۱ - \frac{DW}{۲} \qquad (۴-۵۸)$$

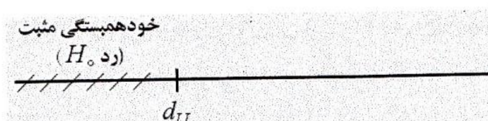
برای nهای نسبتاً بزرگ، h دارای توزیع نرمال با میانگین ۰ و واریانس ۱ می‌باشد. لذا توزیع h، نرمال استاندارد خواهد بود. برای هر توزیعِ نرمال استاندارد رابطه زیر برقرار است:

$$P(-۱/۹۶ \leq h \leq ۱/۹۶) = ۰/۹۵ \qquad (۴-۵۹)$$

بنابراین اگر h بین ۱/۹۶- و ۱/۹۶ قرار بگیرد در این صورت به معنی «عدم خودهمبستگی» است. اگر h کوچکتر از ۱/۹۶- باشد بیانگر خودهمبستگی منفی و اگر بزرگتر از ۱/۹۶ باشد به معنی خودهمبستگیِ مثبت است.

مشکل دیگر در خصوص آزمون DW این است که در برخی موارد، مقدار DW در ناحیه عدم تصمیم‌گیری قرار می‌گیرد. در چنین شرایطی ثابت می‌شود که تقریباً d_U مقدار بحرانی جهت معنی‌دار بودن آزمون خودهمبستگی می‌باشد. حالت‌های مختلفی وجود دارد که عبارتند از:

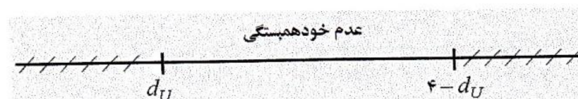
الف) برای آزمون $\rho = ۰$ در مقابل $\rho > ۰$ اگر $DW < d_U$ باشد، در این صورت $\rho = ۰$ در سطح احتمال α رد می‌شود و معادله دارای خودهمبستگی مثبت است.



ب) برای آزمون $\rho = ۰$ در مقابل $\rho < ۰$ اگر $DW > ۴ - d_U$ یا $۴ - DW < d_U$ باشد، در این صورت $\rho = ۰$ در سطح احتمال α رد می‌شود و خودهمبستگی منفی وجود دارد.



ج) برای آزمون $\rho = 0$ در مقابل $\rho \neq 0$ اگر $DW < d_U$ یا $4 - DW < d_U$ باشد، در این صوت $\rho = 0$ در سطح احتمال 2α رد می‌شود و معادله دارای خودهمبستگی مثبت یا منفی است[1].



آزمون دوربین- واتسون در Eviews

وقتی معادله‌ای با Eviews برآورد می‌شود، همراه با نتایج آن، مقدار آماره DW نیز تحت عنوان Durbin-Watson Stat ارائه می‌دهد. با داشتن مقدار آماره DW و مقایسه آن با مقادیر بحرانی که از جدول مربوطه به دست می‌آید، می‌توان راجع به وجود خودهمبستگی قضاوت نمود.

۴-۴-۴ آزمون بروش-گادفری[2]

به‌خاطر داریم که روش دوربین- واتسون برای آزمون خودهمبستگی مرتبه اول است. بنابراین آزمون DW فقط برای شرایط خاص خود قابل کاربرد می‌باشد و اشکال دیگر خودهمبستگی را آزمون نمی‌کند.

به‌عنوان مثال اگر همبستگی بین e_t و e_{t-1} وجود نداشته ولی همبستگی بین e_t و e_{t-2} وجود داشته باشد، در این صورت DW نمی‌تواند هیچ توضیحی راجع به این نوع از خودهمبستگی ارائه نماید. یک راه این است که در محاسبه DW به‌جای e_{t-1} از e_{t-2} استفاده کنیم. در این صورت ضریب همبستگی بین e_t و e_{t-2} محاسبه می‌شود. اما در عمل در نرم‌افزارهای مرسوم، چنین مواردی در نظر گرفته نشده است و همچنین جداولی برای مقادیر بحرانی آن محاسبه نشده است.

برای حل مشکل فوق یک آزمون عمومی در نظر گرفته شده که می‌تواند خودهمبستگی مرتبه rام را آزمون کند. بدین منظور معادله زیر که خودهمبستگی مرتبه rام را نشان می‌دهد، در نظر بگیرید:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_r u_{t-r} + v_t \ , \ v_t \sim N(0, \sigma_v^2) \qquad (4-60)$$

فرضیه‌های مورد نظر عبارتند از:

$$H_0 : \rho_1 = 0 \ , \ \rho_2 = 0 \ , \ \ldots \ , \ \rho_r = 0 \quad \text{عدم‌خودهمبستگی} \qquad (4-61)$$

$$H_1 : \rho_1 \neq 0 \ \text{یا} \ \rho_2 \neq 0 \ \text{یا} \ \ldots \ \text{یا} \ \rho_r \neq 0 \quad \text{وجود خودهمبستگی}$$

[1] Gujarati(2004), P.471
[2] Breusch-Godfrey

تحت فرضیه H_0، خطاهای سال جاری هیچ ارتباطی با r مقادیر قبلی خود ندارند. مراحل انجام این آزمون عبارت است از:

۱- معادله مورد نظر را با روش OLS برآورد می کنیم:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

۲- e_t را حساب کرده و آن را روی متغیرهای توضیحی و خطاهای تأخیری برازش کرده و R^2 آن را حساب می کنیم:

$$e_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \rho_1 e_{t-1} + \cdots + \rho_r e_{t-r} + v_t \qquad (۴-۶۲)$$

اگر تعداد مشاهدات برابر با n باشد، در این صورت می توان برای آزمون این فرضیه از χ^2 استفاده نمود:

$$(n-r)R^2 \sim \chi^2_r \qquad (۴-۶۳)$$

توجه شود که در اینجا تعداد مشاهدات برای برآورد معادله e_t فقط $n-r$ می‌باشد. اگر مقدار آماره $(n-r)R^2$ بزرگتر از مقدار بحرانی (که از جدول به دست می‌آید) باشد، در این صورت فرضیه H_0 رد می‌شود و معادله رگرسیون دارای خودهمبستگی است. هر چند آزمون بروش-گادفری عمومی‌تر از آزمون DW است ولی مشکل آن در تعیین r، یعنی تعداد وقفه‌ها، است.

آزمون بروش- گادفری در Eviews **فایل data2**

برای انجام آزمون بروش-گادفری جهت شناسایی خودهمبستگی، ابتدا معادله مورد نظر را تخمین می‌زنیم. سپس در پنجره‌ای که نتایج تخمین نشان داده شده است (پنجره Equation) مسیر زیر را انتخاب می‌کنیم:

View → Residual Diagnostic → Serial Correlation LM Test

در این حالت پنجره Lag Specification باز می‌شود:



که بایستی تعداد وقفه‌ها (r) را به آن بدهیم. با انتخاب OK، معادله e_t یعنی معادله ۲۲-٤ برآورد می‌شود.

Equation: UNTITLED   Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.641424	Prob. F(2,30)	0.5336
Obs*R-squared	1.435281	Prob. Chi-Square(2)	0.4879

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 01/25/11   Time: 05:18
Sample: 1346 1380
Included observations: 35
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.110183	1.001510	0.110017	0.9131
GL	-0.017235	0.186703	-0.092311	0.9271
GK	-0.009106	0.093942	-0.096937	0.9234
RESID(-1)	0.209243	0.186655	1.121017	0.2712
RESID(-2)	-0.054389	0.201551	-0.269852	0.7891

R-squared	0.041008	Mean dependent var	1.32E-15
Adjusted R-squared	-0.086858	S.D. dependent var	4.129714
S.E. of regression	4.305328	Akaike info criterion	5.889148
Sum squared resid	556.0756	Schwarz criterion	6.111340
Log likelihood	-98.06008	Hannan-Quinn criter.	5.965848
F-statistic	0.320712	Durbin-Watson stat	1.925825
Prob(F-statistic)	0.861846		

در اینجا نیز مجدداً مقادیر F و $(n-r)R^۲$ با احتمال‌های مربوطه نشان می‌دهد که ایـن مـدل خودهمبـستگی نـدارد، مقـدار F و همچنین $\chi^۲=(n-r)R^۲=۳۵(۰/۰۴۱)=۱/۴۳$ کوچک هستند و در ناحیه بحرانی قرار ندارند و لذا فرضیه $H_۰$ رد نمی‌شود.

٥-٤-٤ پیامد نادیده گرفتن خودهمبستگی

اگر خودهمبستگی وجود داشته باشد ولی آن را نادیده بگیریم، تخمین ضرایب بـدون تـورش ولی ناکارآ می‌باشند. این ناکارایی حتی در نمونه‌های بزرگ نیز از بین نمی‌رود و انحراف معیارهـا دارای خطا می‌باشند[1]. این امر منجربه استنتاج‌های نادرست می‌شود. در حـالتی کـه خودهمبـستگی

۱- به‌عنوان مثال در $Y_t = \alpha + \beta_t X_t + u_t$ که u_t خود همبستگی مرتبه اول به‌شکل $u_t = \rho u_{t-۱} + \varepsilon_t$ دارد، در ایـن حالت تخمین زنندهٔ OLS برای β که به صورت $\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^۲}$ است، دارای واریانس زیر می‌باشد:

$$\text{var}(\hat{\beta}) = \frac{\sigma^۲}{\sum x_t^۲} + ۲\frac{\sigma^۲}{(\sum x_t^۲)^۲}\left(\rho\sum_{t=۱}^{n-۱} x_t x_{t+۱} + \rho^۲\sum_{t=۱}^{n-۲} x_t x_{t+۲} + \cdots + \rho^{n-۱} x_۱ x_n\right)$$

مثبت وجود داشته باشد انحراف معیارها کمتر از حد واقعی برآورد می‌شوند. بدین معنی که در این حالت، روش OLS انحراف معیار را کم برآورد می‌کند. در نتیجه موجب افزایش tها شده و ضرایب برآورد شده گرایش به معنادار شدن دارند و لذا مرتکب خطای نوع اول (رد H_0 در حالی که H_0 یک فرضیه درست است) می‌شویم. علاوه بر این، اگر خودهمبستگیِ مثبت وجود داشته باشد و آن را نادیده بگیریم به احتمال زیاد منجربه افزایش مقدار R^2 می‌شود زیرا خودهمبستگی، مقدار واریانس جمله خطا را کم برآورد می‌کند.

۶-٤-٤ تخمین ضرایب در حالت خودهمبستگی

اگر شکل خودهمبستگی را بدانیم می‌توان بـه سادگی از روش OLS استفاده نمود. یکی از روش‌های شناخته شده برای تخمین ضرایب در حالت خودهمبستگی، روش کوکران- اورکات[1] می‌باشد. این روش معمولاً برای خودهمبستگی مرتبه اول طرح شده است. بدین منظور مدل زیر را در نظر بگیرید که در آن u_t دارای خودهمبستگی مرتبه اول است:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (۴-۶۴)$$

$$u_t = \rho u_{t-1} + v_t \qquad (۴-۶۵)$$

فرض کنید که u_t دارای میانگین صفر باشد. مدل فوق را با یک دوره تأخیر نوشته و در ρ ضرب می‌کنیم:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \qquad (۴-۶۶)$$

حال (۶۶-۴) را از (۶۴-۴) کم می‌کنیم:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + (u_t - \rho u_{t-1}) \qquad (۴-۶۷)$$

در (۶۷-۴) به‌جای پرانتز آخر از رابطه (۶۵-۴) قرار می‌دهیم:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + v_t \qquad (۴-۶۸)$$

با تعریف $Y_t - \rho Y_{t-1} = Y_t^*$، $\beta_1^* = \beta_1(1-\rho)$ و $X_2 - \rho X_{t-1} = X_t^*$، معادله (۶۸-۴) را به‌صورت زیر می‌نویسیم:

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + v_t \qquad (۴-۶۹)$$

---

[1] Cocran-Orkat

از آنجا که مدل (۶۹-۴) خودهمبستگی ندارد، لذا می‌توان آن را با روش OLS تخمین زد. اما مدل (۶۹-۴) مبتنی بر داده‌های جدیدی است که به‌صورت Y_t^* و X_t^* هستند. محاسبهٔ چنین داده‌هایی نیاز به مقدار ρ دارد. در عمل مقدار ρ معلوم نیست. یکی از روش‌های ساده این است که ابتدا با تخمین مدل (۶۴-۴) مقدار DW را به‌دست آورده و با استفاده از رابطه $DW = 2(1-\hat{\rho})$، مقدار $\hat{\rho}$ را حساب نمود. سپس بر اساس $\hat{\rho}$، مقادیر جدید یعنی X_t^* و Y_t^* را محاسبه نموده و آنگاه مدل (۶۹-۴) را با روش OLS برآورد نمائیم.

روش تکراری کوکران-اورکات یکی از روش‌های متداول است که مراحل آن به‌صورت زیر است:

۱- ابتدا معادله $Y_t = \beta_1 + \beta_2 X_t + u_t$ را با روش OLS برآورد کرده و باقیمانده‌ها ($\hat{u}_t$ یا e_t) را حساب می‌کنیم.

۲- بر اساس باقیمانده‌های مرحله اول، مدل $\hat{u}_t = \rho\hat{u}_{t-1} + v_t$ یا $e_t = \rho e_{t-1} + v_t$ را برآورد می‌کنیم که تخمین $\hat{\rho}$ به دست می‌آید:

$$\hat{\rho} = \frac{\sum e_{t-1}e_t}{\sum e_{t-1}^2} \qquad (۴-۷۰)$$

۳- با استفاده از $\hat{\rho}$ مقادیر X_t^* و Y_t^* را حساب کرده و سپس روش OLS را برای تخمین مدل (۳۹-۴) به‌کار می‌بریم.

۴- تکرار مراحل ۲ و ۳: در اینجا ابتدا در مرحله ۳ مقادیر خطاها را حساب کرده و مجدداً مرحله ۲ را تکرار می‌کنیم تا به تخمین جدیدی از $\hat{\rho}$ برسیم و سپس با $\hat{\rho}$ جدید به مقادیر جدیدی برای X_t^* و Y_t^* می‌رسیم و بر اساس آنها مجدداً مدل (۵۹-۴) را برآورد می‌کنیم. این تکرارها تا جایی ادامه می‌یابد که مقدار $\hat{\rho}$ تقریباً ثابت بماند و با تکرار مرحلهٔ بعدی، تغییر محسوسی در آن ایجاد نشود. در چنین شرایطی معمولاً عدد کوچکی را برای همگرایی $\hat{\rho}$ در نظر می‌گیرند. مثلاً اگر از یک مرحله به مرحله دیگر تغییرات $\hat{\rho}$ کمتر از ۰/۰۰۱ باشد، در این صورت $\hat{\rho}$ تقریباً به مقدار ثابتی رسیده است. اگر مقدار $\hat{\rho}$ در مرحلهٔ m را با $\hat{\rho}_m$ و در مرحلهٔ $m+1$ را با $\hat{\rho}_{m+1}$ نشان دهیم، در این صورت هدف این است که تکرارها تا جایی ادامه یابد که شرط $|\hat{\rho}_m - \hat{\rho}_{m+1}| < 0/001$ تأمین شود.

نگاه دیگر به موضوع خودهمبستگی این است که آن را به‌جای یک «مشکل»، یک «فرصت» می‌داند. بدین معنی که «وجود خودهمبستگی» به معنی یک «اطلاعات مهم» است و این نشان از تصریح نامناسب مدل دارد. وقتی $\hat{u}_t$ها همبستگی داشته باشند، بدان معنا است که مدل‌سازی Y_t مناسب نبوده است و ساختار پویای آن را در نظر نگرفته‌ایم. در واقع در داخل u_tها عامل یا متغیری وجود دارد که باعث همبستگی سریالی بین آنها شده است و لذا به‌جای رفع خودهمبستگی، باید مدل‌سازی Y_t را تغییر داده و آن را پویا نمائیم.

۷-٤-٤ خودهمبستگی و مدل‌های پویا

مدل‌ها که تاکنون بررسی نمودیم مدل‌های ایستا هستند. به‌عنوان مثال مدل ایستای یک متغیره عبارت است از:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (۴-۷۱)$$

مدل (۴-۷۱) بیانگر یک رابطه ایستا و آنی بین متغیرهاست به‌گونه‌ای که اگر X در زمان t تغییر کند منجر به تغییر آنی در Y خواهد شد. اما این تحلیل را می‌توان به‌صورت پویا بیان نمود و مدل (۴-۷۱) را به‌صورت زیر نوشت که مشابه مدل (۴-۶۸) می‌باشد:

$$Y_t = \beta_1 + \beta_2 X_t + \gamma_1 X_{t-1} + \gamma_2 Y_{t-1} + u_t \qquad (۴-۷۲)$$

در مدل (۴-۷۲) علاوه بر رابطهٔ بین مقدار جاری متغیرها، بین مقدار آنها در طی زمان‌های مختلف نیز ارتباط برقرار شده است. به‌گونه‌ای که اگر X در سال t-۱ تغییر کند، هم بر Y در سال t و هم بر Y در سال $t-1$ اثر خواهد گذاشت. برای بررسی این اثرات، می‌توان معادله (۴-۷۲) را به‌صورت زیر نوشت:

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + \gamma_1 X_{t-2} + \gamma_2 Y_{t-2} + u_{t-1} \qquad (۴-۷۳)$$

$$Y_t = \beta_1 + \beta_2 X_t + \gamma_1 X_{t-1} + \gamma_2 Y_{t-1} + u_t \qquad (۴-۷۴)$$

$$Y_{t+1} = \beta_1 + \beta_2 X_{t+1} + \gamma_1 X_t + \gamma_2 Y_t + u_{t+1} \qquad (۴-۷۵)$$

$$Y_{t+2} = \beta_1 + \beta_2 X_{t+2} + \gamma_1 X_{t+1} + \gamma_2 Y_{t+1} + u_{t+2} \qquad (۴-۷۶)$$

اگر X_{t-1} تغییر کند طبق معادله (۴-۷۳) بر Y_{t-1} اثر می‌گذارد. از طرف دیگر طبق معادله (۴-۷۴) هم X_{t-1} و هم Y_{t-1} بر Y_t اثر می‌گذارند. همچنین طبق معادله (۴-۷۵)، تغییرات Y_t بر Y_{t+1} اثر می‌گذارد. این تغییرات همچنان ادامه می‌یابد.

مدل‌هایی که فقط شامل متغیرهای توضیحیِ تأخیری باشند موسوم به مدل‌های باوقفه توزیعی[1] هستند. اما مدل‌ها که در آنها هم متغیر توضیحی و هم متغیر وابسته دارای وقفه باشند، موسوم به مدل‌های باوقفه توزیعی خودرگرسیونی (ARDL)[2] می‌باشند که در فصل دهم بررسی شده‌اند.

حال سؤال این است که چرا باید وقفه‌ها را وارد یک مدل کنیم؟

مقادیر باوقفهٔ متغیرهای توضیحی یا متغیر وابسته به معنی این است که ساختار مدل، پویا می‌باشد. یکی از دلایل مهمی که در این زمینه وجود دارد، کُندی تعدیل متغیر وابسته است. اغلب اوقات تغییرات متغیر توضیحی اثر خود را به‌طور آنی بر متغیر وابسته نمی‌گذارد، بلکه اثرات آن نیاز به گذشت زمان دارد. به‌عنوان مثال تغییر در سیاست دولت ممکن است نیاز به گذشت چند ماه یا سال داشته باشد تا بتواند اثرات خود را بر متغیرهای اقتصادی بگذارد. همچنین در بازارهای مالی مانند بورس وقتی تغییری رخ می‌دهد و به‌صورت اطلاعات جدید دریافت می‌شود، ممکن است افراد به‌طور آنی به آن واکنش نشان ندهند، بلکه به‌صورت تدریجی خود را با آن تطبیق دهند. سرعت و مقدار واکنش آنها همچنین بستگی به این دارد که آیا آنها تغییرات را دائمی می‌دانند یا موقتی. همچنین تأخیر در واکنش ممکن است ناشی از عوامل نهادی و فنی باشد. به‌عنوان مثال وجود فناوری‌های پیشرفته می‌تواند به سرعت خرید و فروش سهام کمک نماید و از این رو واکنش‌ها را سریع‌تر می‌نماید.

خودهمبستگی و مدل‌های پویا در Eviews فایل data1

در Eviews می‌توان متغیرهای X_t، X_{t-1}، X_{t-2}، Y_{t-1} را به‌عنوان متغیر توضیحی به مدل مورد نظر اضافه نمود. در این حالت آنها را به‌صورت X، X(-1)، X(-2) و Y(-1) وارد معادله می‌کنیم.

اگر سرمایه‌گذاری خصوصی(IP) را روی نرخ بهره حقیقی (RR) برازش کنیم، این معادله با اجرای فرمان LS IP C RR برآورد می‌شود که نتایج آن عبارت است از:

[1] Distributed lag
[2] Autoregressive Distributed Lag

Equation: UNTITLED Workfile: S3::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: IP
Method: Least Squares
Date: 01/25/11 Time: 06:00
Sample: 1345 1375
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20571.71	2295.966	8.959940	0.0000
RR	-427.0836	209.5819	-2.037789	0.0508

R-squared	0.125257	Mean dependent var	23205.85
Adjusted R-squared	0.095093	S.D. dependent var	11106.10
S.E. of regression	10564.86	Akaike info criterion	21.43079
Sum squared resid	3.24E+09	Schwarz criterion	21.52331
Log likelihood	-330.1773	Hannan-Quinn criter.	21.46095
F-statistic	4.152584	Durbin-Watson stat	0.812045
Prob(F-statistic)	0.050785		

از آنجا که $DW=۰/۸۱$ می‌باشد، لذا این معادله دارای خودهمبستگی است. همان‌طور که اشاره شد، خودهمبستگی یکی از علائم نادرست بودن مدل است. حال مدل را به‌صورت پویا در نظر گرفته و متغیر وابسته تأخیری را به آن اضافه مـی‌کنـیم، بـرای برآورد این مدل فرمان زیر را اجرا می‌کنیم:

```
LS    IP    C    RR  IP(-1)
```

نتیجه حاصله عبارت است از:

Equation: UNTITLED Workfile: S3::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: IP
Method: Least Squares
Date: 01/25/11 Time: 06:07
Sample (adjusted): 1346 1375
Included observations: 30 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7229.362	3132.899	2.307563	0.0289
RR	41.49133	168.3648	0.246437	0.8072
IP(-1)	0.736791	0.137058	5.375770	0.0000

R-squared	0.561056	Mean dependent var	23816.94
Adjusted R-squared	0.528541	S.D. dependent var	10752.83
S.E. of regression	7383.198	Akaike info criterion	20.74644
Sum squared resid	1.47E+09	Schwarz criterion	20.88656
Log likelihood	-308.1966	Hannan-Quinn criter.	20.79127
F-statistic	17.25560	Durbin-Watson stat	2.229264
Prob(F-statistic)	0.000015		

برای بررسی خودهمبستگی از h دوربین استفاده می‌کنیم، زیرا این مدل دارای متغیر وابسته تأخیری است:

$$h=\left(۱-\frac{DW}{۲}\right)\sqrt{\frac{n}{۱-n\,var(\hat{\theta})}}=(۱-\frac{۲/۲۳}{۲})\sqrt{\frac{۳۰}{۱-(۳۰)(۰/۱۳۷۱)^{۲}}}=-۰/۹۷۶۵$$

چون مقدار ۰/۹۷٦۵- در فاصله (۱/۹٦ و ۱/۹٦-) قرار دارد، لذا وجود خودهمبستگی رد می‌شود.

۸-٤-٤ خودهمبستگی و فرایندهای خودرگرسیون (AR)[1] و میانگین متحرک (MA)[2]

فرایند خودهمبستگی مرتبه اول که به شکل $u_t = \rho u_{t-1} + \varepsilon_t$ است را AR(۱) می‌گویند که معروف به خودرگرسیون مرتبه اول است. همچنین $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$ را فرایند AR(۲) می‌گویند که بیانگر خودهمبستگی مرتبه دوم است. اما نوع دیگری از این فرایندها را می‌توان تعریف کرد که معروف به میانگین متحرک (MA) هستند. به‌عنوان مثال اگر خودهمبستگی به‌شکل $u_t = \varepsilon_t + \theta\varepsilon_{t-1}$ باشد آن را میانگین متحرک مرتبه اول می‌گویند که با MA(۱) نشان داده می‌شود. اینکه چگونه فرایند MA بیانگر خودهمبستگی است، می‌توان آن را به AR تبدیل نمود. همچنین می‌توان فرایند AR نیز تحت شرایطی، تبدیل به MA نمود. قابلیت تبدیل این فرایندها و جزئیات آن در فصل یازدهم بحث شده است. به‌عنوان مثال می‌توان AR(۱) را به‌صورت $u_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots$ نوشت که بیانگر MA(∞) است. علاوه بر نکات فوق، در معادله $u_t = \rho u_{t-1} + \varepsilon_t$، اگر $\rho = -1$ باشد، آنگاه معادله (۴-۶۷) را به‌صورت زیر می‌توان نوشت:

$$Y_t + Y_{t-1} = 2\beta_1 + \beta_2(X_t + X_{t-1}) + (u_t + u_{t-1})$$

با تقسیم طرفین بر ۲ خواهیم داشت:

$$\frac{Y_t + Y_{t-1}}{2} = \beta_1 + \beta_2\frac{X_t + X_{t-1}}{2} + \frac{u_t + u_{t-1}}{2} \qquad (۴-۴۷)$$

معادله فوق بیانگر میانگین متحرک متغیرها است.

به‌هرحال، نتیجه کلی بحث فوق آن است که اگر معادله رگرسیون دارای خودهمبستگی باشد، ممکن است از یکی از فرایندهای AR یا MA تبعیت کند و لذا برای حل مشکل خودهمبستگی می‌توان فرایندهای AR یا MA را در نظر گرفت.

رفع خودهمبستگی با استفاده از AR و MA در Eviews

در Eviews می‌توان معادله رگرسیون را همراه با AR و MA تخمین زد. به‌عنوان مثال برای رفع خودهمبستگی مرتبه اول می‌توان فرایند AR(1) را در نظر گرفت و با فرمان LS Y C X AR(1) آن را تخمین زد. همچنین برای MA(1) از فرمان LS Y C X MA(1) استفاده می‌شود. علاوه بر این، می‌توان هر دو را با فرمان LS Y C X AR(1) MA(1) برآورد نمود.

[1] Autoregressive
[2] Moving Average

۵-٤ فرض ٤: غیر تصادفی بودن متغیرهای توضیحی

غیرتصادفی بودن متغیرهای تصادفی (X_t ها) یکی از فروض کلاسیک است که موجب می‌شود تا تخمین‌زننده‌های OLS بدون تورش باشند. غیرتصادفی بودن X_t موجب استقلال آن از متغیر تصادفی u_t می‌گردد. اما اگر X_t تصادفی باشد، تخمین‌زننده‌های OLS مجدداً سازگار و بدون تورش خواهند بود. بدین منظور تخمین‌زننده OLS را در نظر بگیرید که عبارت است از:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}(\mathbf{X'y}) = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'u} \qquad (۴-۷۸)$$

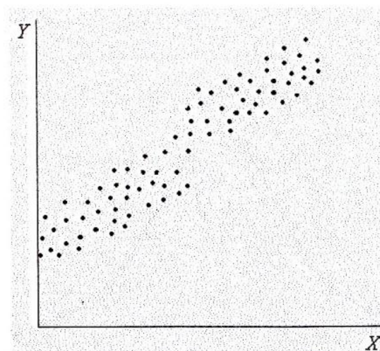
امید ریاضی $\hat{\boldsymbol{\beta}}$ عبارت است از:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}E(\mathbf{X'u}) \qquad (۴-۷۹)$$

برای آنکه $\hat{\boldsymbol{\beta}}$ بدون تورش باشد، لازم است که $\mathbf{X}$ و $\mathbf{u}$ مستقل باشند، هرچند که $\mathbf{X}$ تصادفی باشد.

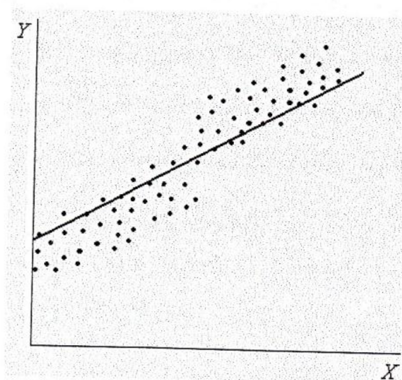
با فرض مستقل بودن $\mathbf{X}$ و $\mathbf{u}$، رابطه $E(\mathbf{X'u})=E(\mathbf{X'})E(\mathbf{u})=\mathbf{o}$ برقرار است و لذا $\hat{\boldsymbol{\beta}}$ بدون تورش خواهد بود. اما مشکل زمانی به‌وجود می‌آید که $\mathbf{X}$ و $\mathbf{u}$ مستقل نباشند. در این صورت $E(\mathbf{X'u}) \neq ۰$ است و لذا $\hat{\boldsymbol{\beta}}$ تخمین‌زنندهٔ بدون تورش نخواهد بود.

همبستگی بین u_t و X_t می‌تولند ناشی از عوامل مختلفی باشد. به‌عنوان مثال اگر مشاهدات Y و X مانند نمودار زیر باشد، در این صورت رگرسیون دچار شکستگی شده است (فصل ششم را ببینید) و بدون توجه به آن، رگرسیون را برآورد کرده‌ایم.



نمودار ۸-٤

با ترسیم خط رگرسیون، وضعیت باقیمانده‌ها روشن‌تر می‌شود.



نمودار ۹-٤: همبستگی مثبت بین u و X

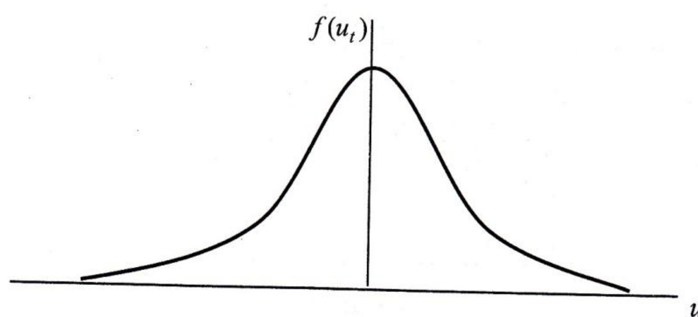
نمودار فوق نشان می‌دهد که با افزایش X مقادیر جملهٔ خطا افزایش می‌یابد، زیرا به‌ازای X_i های کوچک، u_iها منفی هستند و به‌ازای X_i های بزرگ، u_iها مثبت می‌باشند. این مثالی از همبستگی مثبت بین u و X است.

٤-٦ فرض ٥: فرض نرمال بودن u_t

یکی از فروض مربوط به جمله خطاها این است که u_t توزیع نرمال دارد.

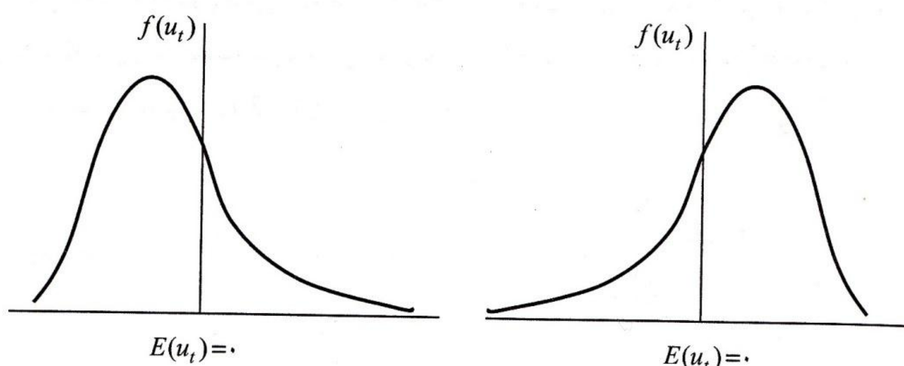
$$u_t \sim N(۰,\sigma^۲) \qquad (۴-۸۰)$$

به‌خاطر داریم که تمام آزمون‌های مربوط به معنادار بودن ضرایب مبتنی بر فرض نرمال بودن u_t است. اگر u_t دارای توزیع نرمال با میانگین صفر باشد، آنگاه در اطراف میانگین خود به‌صورت متقارن توزیع می‌شود.



نمودار ۱۰-٤: توزیع متقارن

اما اگر u_t نرمال نباشد، ممکن است قرینه نباشد و دارای چولگی باشد.



نمودار ۱۱-٤: توزیع‌های نامتقارن (چوله)

از طرف دیگر $\frac{u_t - E(u_t)}{\sigma} = \frac{u_t}{\sigma}$ توزیع نرمال استاندارد با میانگین صفر و واریانس ۱ دارد. هر توزیع نرمال استاندارد اولاً قرینه است، یعنی ضریب چولگی آن صفر است و ثانیاً ضریب کشیدگی آن برابر با ۳ می‌باشد. بنابراین، برای بررسی انحراف از توزیع نرمال، بایستی ضرایب چولگی و کشیدگی را محاسبه و بررسی نماییم.

جارکیو و برا[1] (۱۹۸۱) ایدهٔ فوق را در قالب یک آزمون ارائه نمودند. برای متغیر تصادفی u با امید ریاضی صفر و واریانس $\sigma^۲$، ضرایب چولگی ($A_۱$) و کشیدگی ($A_۲$) عبارتند از:

$$A_۱ = \frac{E(u_t^۳)}{\sigma^۳} \quad , \quad A_۲ = \frac{E(u_t^۴)}{\sigma^۴} \qquad (۴-۵۱)$$

اگر u_t قرینه باشد، آنگاه $A_۱$ تقریباً صفر است. همچنین اگر $A_۲$ نزدیک به ۳ باشد در این صورت کشیدگی u_t به نرمال استاندارد نزدیک است و $A_۲ - ۳$ نیز تقریباً صفر خواهد بود. بنابراین،اگر u_t نرمال باشد $A_۱$ و $A_۲ - ۳$ تقریباً صفر خواهند بود. بر این اساس، آماره جارکیو- برا برای آزمون نرمال بودن عبارت است از:

$$JB = n\left(\frac{A_۱^۲}{۶} + \frac{(A_۲ - ۳)^۲}{۲۴}\right) \qquad (۴-۵۲)$$

[1] Jarque and Bera

تابع JB تحت فرضیه H_0 (فرضیه‌ای که طبق آن، u نرمال است) بـه توزیـع χ^2 بـا درجـه آزادی ۲ گرایش دارد که مقدار بحرانی آن برابر با $\chi^2_{0.05,2} = 5/99$ است. برای محاسبه JB می‌تـوان A_1 و A_2 را با استفاده از e_tها محاسبه نمود. بدین منظور بایستی $E(u_t^3)$ و $E(u_t^4)$ و همچنین $E(u_t^2) = \sigma^2$ را برآورد نماییم. تخمین $E(u_t^3)$ و $E(u_t^4)$ برابر است با:

$$\hat{E}(u_t^3) = \frac{\sum_{t=1}^{n} e_t^3}{n} \qquad \hat{E}(u_t^4) = \frac{\sum_{t=1}^{n} e_t^4}{n} \tag{4-53}$$

اگر A_1 یا A_2 و یا هر دو آنها بزرگ باشند، آنگاه مقدار χ^2 را افزایش داده و در ناحیه بحرانی قرار می‌گیرد که فرضیه H_0 (نرمال بودن e_t) را رد می‌کند و بدین معنی است که u_t از نظر چولگی یا کشیدگی، اختلاف زیادی با توزیع نرمال دارد.

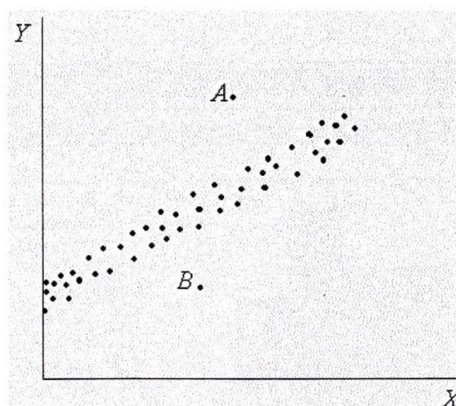
آزمون نرمال بودن با استفاده از Eviews **فایل** data2

برای آزمون نرمال بودن ابتدا معادله اصلی را برآورد کرده و سپس در پنجره‌ای که نتایج تخمین را نشان می‌دهد، به‌صورت زیر عمل می‌کنیم:

View → Residual Diagnostics → Histogram–Normality Test

اگر باقیمانده‌ها به‌صورت نرمال توزیع شده باشند، نمودار آن تقریباً به‌صورت نرمال است و آماره JB معنی‌دار نمی‌باشد. این بدان معنا است که در چنین شرایطی مقدار آمارهٔ JB کوچک می‌باشد و مقدار احتمال که در زیر آماره JB داده می‌شود، بزرگتر از ۰/۰۵ است. در واقع بزرگ بودن احتمال معادل با کوچک بودن آمارهٔ JB است که در این صورت فرضیه نرمال بودن رد نمی‌شود. بر عکس اگر مقدار احتمال کوچک باشد مقدار آمارهٔ JB بزرگ خواهد بود و لذا فرضیه نرمال بودن رد می‌شود. به‌طور کلی اگر JB>۵/٦٦ باشد، نرمال بودن رد می‌شود، در این صورت مقدار احتمال که در زیر آماره JB داده می‌شود کوچکتر از ۰/۰۵ خواهد بود.

به‌عنوان مثال در رگرسیون نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) روی رشد نیروی کار (GL) و رشد سرمایه (GK)، نتایج آن در پنجره Equation نشان داده می‌شود. در این پنجره از منوی View آزمون نرمال بودن را بررسی می‌کنیم که نتایج عبارت است از:



نمودار ۱۲-٤: مشاهدات دور افتاده

وجود مشاهدات دورافتاده موجب کشیدگی بیش از حد می‌شود و لذا آن را تبدیل به یک توزیع دم‌پهن می‌کند. این پدیده‌های است که در داده‌های بورس معمولاً مشاهده می‌شود.

مشاهدات دورافتاده به‌دلیل وقوع حوادث نادر[1] یا به هر دلیل دیگری به وجود می‌آید. یکی از راه‌حل‌های این پدیده، استفاده از متغیرهای مجازی[2] است. جزئیات این بحث در فصل پنجم ارائه شده است. کارکرد متغیرهای مجازی بدین صورت است که می‌تواند اثر مشاهدات دورافتاده را به‌صورت جداگانه لحاظ کند و تأثیر آنها را بر باقیمانده‌ها (e_t) حذف نماید. به‌عنوان مثال فرض کنید که دوره ۸۰-۱۳۵۰ را بررسی می‌کنیم. در یکی از این سالها مثلاً در سال ۱۳۵۷ مقدار متغیر مورد نظر تفاوت فاحشی با بقیه سال‌ها دارد. در اینجا یک متغیر جدید (مثلاً به نام D57) معرفی می‌کنیم که مقدار آن برای سال ۱۳۵۷ برابر با ۱ و برای بقیه سال‌ها برابر با صفر است. در این حالت در مدل جدید، اثر مشاهدهٔ سال ۱۳۵۷ بر روی باقیمانده‌ها تعدیل می‌شود و موجب می‌شود تا باقیمانده‌ها دارای توزیع نرمال باشند.

مسائل

۱- ثابت کنید اگر رگرسیونی بدون عرض از مبدأ برآورد شود، مجموع پسماندها لزوماً صفر نخواهد بود. اگر $R^2 = 1 - \frac{e'e}{Y'Y - n\bar{Y}^2} = 1 - \frac{\sum e_t^2}{\sum y_t^2}$ را حساب کنیم، ممکن است منفی شود.

---

[1] rare events
[2] dummy variables

۲- نمونه ۵تایی زیر را داریم:

X	٤	۲	٥	۸	۲
Y	٦	۳	۱۲	۱٥	٤

با فرض اینکه واریانس ناهمسانی از نوع $\sigma_t^2 = \sigma^2 X_t^2$ باشد، برآوردهای α، β و انحراف معیار معیار آنها را با روش GLS به‌دست آورید.

۳- مواردی را که آزمون دوربین- واتسون قابل کاربرد نیست نام برده و توضیح دهید.

۴- برای بررسی واریانس ناهمسانی، معادله زیر معرفی شده است. آیا می‌توان پارامترها را به روش OLS تخمین زد؟

$$|u_t| = \sqrt{\alpha + \beta X_t} + \varepsilon_t$$

اگر جواب منفی است آیا روشی برای تخمین این پارامترها وجود دارد؟

۵- گفته می‌شود که مدل‌های لگاریتمی توانایی کاهش مسئله واریانس ناهمسانی را دارند. برای مثال مدل $Y_t = aX_t^{\beta} u_t$ زیر را در نظر بگیرید که می‌توان آن را به‌صورت زیر نوشت:

$$\ln Y_t = \alpha + \beta \ln X_t + \ln u_t \quad , \quad \alpha = \ln a$$

الف) اگر بخواهیم امید ریاضی $\ln u_t$ صفر باشد، u_t چه توزیعی باید داشته باشد؟

ب) اگر $E(u_t) = 1$ باشد، آیا $E(\ln u_t)$ می‌تواند صفر باشد؟

ج) اگر $E(\ln u_t)$ صفر نباشد برای آنکه صفر شود چه می‌توان کرد؟

۶- معادله رگرسیون زیر را در نظر بگیرید:

$$\hat{Y}_t = 2/55 + 0/65 Y_{t-1} + 3/42 X_t \quad ; \quad n = 50 \quad , \quad DW = 1/65$$
$$(2/2) \quad (5/5) \quad (3/1)$$

اعداد داخل پرانتز مقادیر t هستند. خودهمبستگی مرتبه اول را آزمون کنید.

۷- مدل زیر را در نظر بگیرید:

$$Y_t = \alpha + \beta X_t + \gamma Z_t + u_t$$

جملهٔ اختلال از فرآیند خودهمبستگی مرتبه دوم پیروی می‌کنند:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + v_t$$

که v_t فروض کلاسیک را تأمین می‌کند. معادله رگرسیون را تبدیل به معادله‌ای کنید که خودهمبستگی نداشته باشد.

۸- اگر در معادله $Y_t = \alpha + \beta X_t + u_t$ جملهٔ خطا به‌صورت $u_t = \gamma X_t^۲ + \varepsilon_t$ باشد،

الف) کدامیک از فروض کلاسیک نقض شده است؟

ب) تحت چه شرایطی تخمین‌زنندهٔ OLS بدون تورش است؟

۹- اگر معادله رگرسیون فاقد عرض از مبدأ باشد، در چه صورتی فرض $E(u_t) = ۰$ نقض می‌شود؟

۱۰- در مسئله۱ فصل اول، آزمون وایت را برای تشخیص واریانس ناهمسانی انجام دهید.

۱۱- در مسئله۱ فصل اول، آیا معادلهٔ $Y_i = \beta_۱ + \beta_۲ X_i + \beta_۳ Z_i + u_i$ خودهمبستگی مرتبهٔ اول دارد یا نه؟ در صورت وجود خودهمبستگی، آن را رفع کنید.

۱۲- در مسئله۱ فصل اول، نرخ رشد Y را روی نرخ رشد X و Z برازش کنید. آیا این معادله، خودهمبستگی دارد؟

۱۳- در مسئله۱ فصل اول، آزمون خودهمبستگی بروش-گادفری را برای معادلهٔ $Y_i = \beta_۱ + \beta_۲ X_i + \beta_۳ Z_i + u_i$ انجام دهید.

۱۴- در مسئله۱ فصل اول، آزمون نرمال بودن باقیمانده‌ها را برای هر یک از معادلات زیر انجام دهید:

الف) $Y_i = \beta_۱ + \beta_۲ X_i + u_i$

ب) $Y_i = \beta_۱ + \beta_۳ Z_i + u_i$

ج) $Y_i = \beta_۱ + \beta_۲ X_i + \beta_۳ Z_i + u_i$

۱۵- مدل زیر را در نظر بگیرید که خودهمبستگی از فرایند $MA(۱)$ تبعیت می‌کند

$$Y_t = \beta Y_{t-۱} + u_t$$
$$u_t = \varepsilon_t + \theta \varepsilon_{t-۱}$$

ثابت کنید که تخمین β از روش OLS دارای تورش است:

$$\text{plim}\hat{\beta}=\beta+\frac{\theta(۱-\beta^{۲})}{۱+\theta^{۲}+۲\theta\beta}=\beta+\frac{\theta}{۱+\frac{(\theta+\beta)^{۲}}{(۱-\beta^{۲})}}$$

۱۶- استفاده از آماره DW مستلزم وجود عرض از مبدأ و عدم وجود متغیر وابسته تأخیری در سمت راست می‌باشد. اگر این دو شرط نقض شود، مدل زیر را خواهیم داشت:

$$Y_t+\beta Y_{t-۱}+u_t$$
$$u_t=\rho u_{t-۱}+\varepsilon_t$$

ثابت کنید که برای آماره DW شرط زیر برقرار است:

$$\text{plim}DW=۲\left(۱-\frac{\rho\beta(\rho+\beta)}{۱+\rho\beta}\right)$$

۱۷- معادله $Y_t=\alpha+\beta X_t+u_t$ دارای واریانس ناهمسانی به شکل $\sigma_t^{۲}=\sigma^{۲}X_t^{\gamma}$ است. برای رفع واریانس ناهمسانی، چگونه این مدل رابرآورد می‌کنید.

۱۹- مدل زیر با $n=۳۰$ مشاهده، برآورد شده است:

$$\hat{Y}_t=۴/۷۵+۰/۸۸X_t+۰/۴۴Y_{t-۱} \qquad R^{۲}=۰/۹۵$$
$$(۲/۴)\quad(۴/۲)\quad(۳/۵) \qquad DW=۲/۲۱$$

آزمون خودهمبستگی را انجام دهید.

۲۰- اگر مدل $Y_t=\alpha+\beta X_t+u_t$ دارای خودهمبستگی مرتبه دو به‌صورت $u_t=\rho_۱u_{t-۱}+\rho_۲u_{t-۲}+\varepsilon_t$ باشد، مدل فوق را تبدیل به مدلی کنید که خودهمبستگی نداشته باشد.

۲۱- در مدل $Y_t=\alpha+\beta X_t+u_t$ اگر خودهمبستگی مرتبه اول وجود داشته باشد، ثابت کنید که ماتریس واریانس $\mathbf{u}$ برابر است با:

$$E(\mathbf{uu}')=\sigma^{۲}\begin{bmatrix}۱ & \rho & \rho^{۲} & \cdots & \rho^{n-۱}\\ \rho & ۱ & \rho & \cdots & \rho^{n-۲}\\ \rho^{۲} & \rho & ۱ & \cdots & \rho^{n-۳}\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ \rho^{n-۱} & \rho^{n-۱} & \rho^{n-۱} & \cdots & \rho^{n-۱}\end{bmatrix}$$

$\mathbf{u}$ بردار ستونی است که عناصر آن u_tها هستند.

۲۲- رگرسیون بدون عرض از مبدأ را به‌صورت $Y_t = \beta X_t + u_t$ در نظر بگیرید. اگر $\sigma_t^2 = \sigma^2 X_t$ باشد:

الف) وضعیت واریانس ناهمسانی در این مدل را توضیح را توضیح دهیم.

ب) مدل مذکور را با روش OLS برآورد کرده و واریانس آن را حساب کنید.

ج) واریانس ناهمسانی را برطرف کرده و سپس آن را برآورد نمایید.

د) نتایج ب و ج را مقایسه کنید.

ضمیمه فصل ٤: آزمون‌های نقض فروض در Stata

آزمون واریانس ناهمسانی در Stata

فایل data3

آزمون‌های واریانس ناهمسانی و بسیاری از آزمون‌های دیگر، با استفاده از باقیمانده‌ها انجام می‌شود. برای محاسبه eها، بعد از تخمین مدل، فرمان زیر را اجرا می‌کنیم:

```
predict e, residuals
```

حال می‌توان خطاها را در مقابل زمان و یا هر متغیر دیگری رسم کرده و خصوصیات آن را بررسی نمود.

الف) آزمون وایت

ابتدا معادله مورد نظر را تخمین می‌زنیم:

```
. reg i rr d.yno l.i

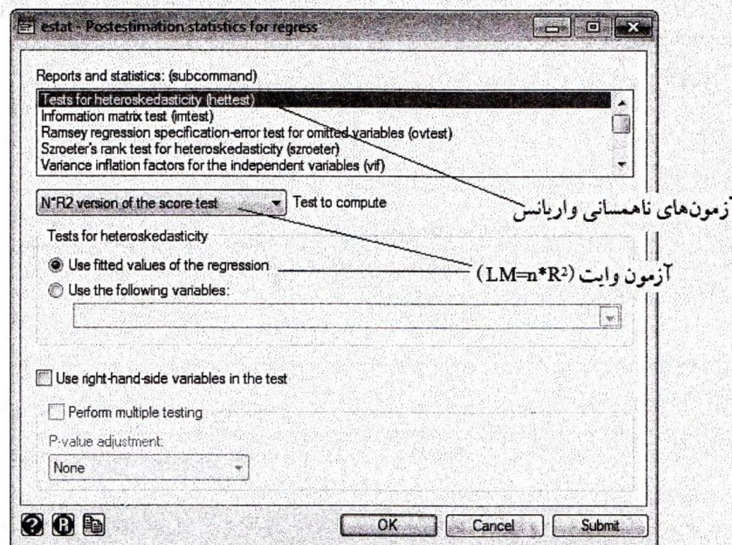
      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  3,    31) =  184.73
       Model |  4.1221e+10     3  1.3740e+10           Prob > F      =  0.0000
    Residual |  2.3058e+09    31  74380117.7           R-squared     =  0.9470
-------------+------------------------------           Adj R-squared =  0.9419
       Total |  4.3527e+10    34  1.2802e+09           Root MSE      =  8624.4

------------------------------------------------------------------------------
           i |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |  -246.8273   200.1467    -1.23   0.227    -655.0293    161.3746
             |
         yno |
         D1. |    1.18792   .1583372     7.50   0.000     .8649894    1.510851
             |
           i |
         L1. |   .8194636   .0523257    15.66   0.000     .7127447    .9261825
             |
       _cons |   5742.107    5095.64     1.13   0.268     -4650.52    16134.73
------------------------------------------------------------------------------
```

حال مسیر زیر را دنبال می‌کنیم:

Statistics → Linear models and related → regression diagnostics → Specification test, ets.

که پنجره زیر باز می شود:



با انتخاب OK نتیجه آزمون وایت عبارت است از:

```
. estat hettest, iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of i

         chi2(1)      =     1.47
         Prob > chi2  =   0.2248
```

مقدار آماره آزمون (یعنی ۱/٤٧) در ناحیه بحرانی قرار ندارد (زیرا احتمال کمتر از ۰/۰۵ است) و لذا همسانی واریانس رد نمی‌شود.

آزمون وایت را می توان با فرمان زیر نیز ا نجام داد:

estat hettest, iid

ب) آزمون بروش- پاگان

آزمون واریانس ناهمسانی بروش- پاگان را با فرمان hettest بعد از تخمین معادله مورد نظر به دست آورد.

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of i

         chi2(1)      =     1.13
         Prob > chi2  =   0.2876
```

مقدار آماره آزمون در ناحیه بحرانی قرار ندارد و لذا فرضیه H_0 رد نمی شود یعنی واریانس همسان است.

برآورد واریانس مستحکم وایت

اگر بخواهیم در معادله رگرسیون، آزمون معنی‌دار بودن ضرایب بر اساس واریانس مستحکم وایت صورت گیرد از فرمان زیر برای برآورد معادله استفاده می‌کنیم:

```
reg y x, vce(robust)
```

به عنوان مثال رگرسیون سرمایه‌گذای روی نرخ بهره و تغییرات درآمد ملی را در نظر بگیرید که قبلاً برآورد گردید:


```
. reg i rr d.yno l.i, vce(robust)

Linear regression                          Number of obs =      35
                                           F(  3,     31) =  442.55
                                           Prob > F      =  0.0000
                                           R-squared     =  0.9470
                                           Root MSE      =  8624.4
```

i	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
rr	-246.8273	248.583	-0.99	0.328	-753.8156	260.1609
yno D1.	1.18792	.1889259	6.29	0.000	.8026033	1.573237
i L1.	.8194636	.0451138	18.16	0.000	.7274534	.9114738
_cons	5742.107	4667.693	1.23	0.228	-3777.716	15261.93

روش GLS در Stata

فرض کنید که عامل واریانس ناهمسانی، متغیر x باشد. مثلاً تابع $\sigma_t^2 = \sigma^2 X_t$ را داشته باشیم. بنابراین در اینجا، وزن‌ها معلوم است و بایستی طرفین معادله را بر $\sqrt{X_t}$ تقسیم کنیم. بدین منظور روش GLS را با فرمان زیر به کار می‌بریم:

```
reg y x [aweight=1/x]
```

توجه شود که Stata، طرفین معادله را بر $\frac{1}{\sqrt{X_t}}$ تقسیم می‌کند.

فرض کنید تابع واریانس به صورت زیر باشد:

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_t + \alpha_2 Z_t$$

برای استفاده از روش GLS بایستی طرفین معادله را بر σ_t تقسیم کنیم. اما چون σ_t را نداریم لذا آن را برآورد می‌کنیم. مراحل انجام این کار و تخمین GLS به صورت زیر است:

۱- معادله را تخمین می‌زنیم

```
reg y x
```

۲- e_t ها را حساب می‌کنیم:

```
predict e, residual
```

۳- لگاریتم e_t^2 را حساب می‌کنیم:

```
gen lne2=log(e^2)
```

٤- رگرسیون lne2 را روی x و z برآورد می‌کنیم:

```
reg lne2 x z
```

٥- مقادیر به دست آمده از این رگرسیون معادل با برآورد σ_t^2 است، لذا آنها را با فرمان زیر حساب می‌کنیم:

```
predict lns2, xb
```

٦- چون $\ln s2=\log(\hat{\sigma}_t^2)$ است، لذا تخمین واریانس را به صورت زیر حساب می‌کنیم:

```
gen wt=exp(lns2)
```

wt معادل با $\hat{\sigma}_t^2$ است.

۷- حال روش GLS را به صورت زیر اعمال می‌کنیم:

```
reg y x, [aweight=1/wt]
```

توجه شود که در روش GLS طرفین معادله را بر جذر wt تقسیم می‌کنیم. در اینجا نیز ما وزن را به صورت معکوس wt تعریف کرده‌ایم ولی Stata آن را به صورت $\frac{1}{\sqrt{wt}}$ در نظر می‌گیرد.

آزمون خود همبستگی در Stata

آزمون‌های خودهمبستگی مبتنی بر استفاده از باقیمانده‌ها هستند. برای محاسبه eها، بعد از تخمین مدل، فرمان زیر را اجرا می‌کنیم:

```
predict e, residuals
```

حال می‌توان خطاها را در مقابل زمان و یا هر متغیر دیگری رسم کرده و خصوصیات آن را بررسی نمود. برای رسم eها در مقابل زمان از فرمان زیر استفاده می‌کنیم:

```
tsline e
```

از طرف دیگر برای بررسی دقیق‌تر، می‌توان خطاهای زمان t را در مقابل زمان t-1 رسم نمود:

```
Twoway (scatter e  l.e)
```

همچنین می‌توان ضریب همبستگی e_t و e_{t-1} را نیز حساب نمود که برآوردی از ρ می‌باشد. بدین منظور، فرمان زیر را اجرا می‌کنیم:

```
correlate e l.e
```

آزمون دوربین- واتسون

بدین منظور، بعد از تخمین معادله موردنظر، فرمان زیر اجرا می‌کنیم:

```
dwstat
```

با اجرای فرمان فوق، مقدار آماره DW به صورت زیر نشان داده می‌شود:

```
. dwstat
Durbin-Watson d-statistic(  2,     21) =  1.177749
```

که DW=۱/۱۷۸ است. ارقام داخل پرانتز بیانگر ۲ پارامتر و ۲۱ مشاهده است.

آزمون نرمال بودن در Stata

آزمون جارک- برا یکی از آزمون‌های نرمال بودن است. مراحل این آزمون عبارت است از:

۱- معادله مورد نظر را تخمین می‌زنیم:

```
reg y x
```

۲- باقیمانده‌ها را حساب می‌کنیم:

```
predict e, residuals
```

۳- شاخص‌های مربوط به e را حساب می‌کنیم:

```
summarize e, detail
```

به جای summarize می‌توان به‌اختصار از sum استفاده کرد.

٤- شاخص‌های مورد نظر را با فرمان return list مشاهده می‌کنیم.
۵- آماره جارک- برا را با فرمان زیر حساب می‌کنیم:

scalar jb=r(N)/6*[r(skewness)^2+0.25*(r(kurtosis)-3)^2]

توجه شود که r(N) حجم مشاهدات است و N بایستی با حروف بزرگ نوشته شود.

```
. scalar jb=r(N)/6*[r(skewness)^2+0.25*(r(kurtosis)-3)^2]
```

٦- مقدار عددی آماره جارک- برا را با فرمان زیر مشاهده می‌کنیم:

di "jarque-bera statistic="jb

```
. di "jarque-bera statistic="jb
jarque-bera statistic=2.6482548
```

۷- اگر jb >۵/۹۹ باشد، نرمال بودن رد می‌شود. در اینجا چون ۵/۹۹>۲/٦٤۸=jb است لذا فرضیه نرمال بودن رد نمی‌شود.

فصل پنجم

متغیرهای مجازی

۱-۵ مقدمه

تاکنون متغیرهای توضیحی را قابل اندازه‌گیری و پیوسته در نظر گرفتیم. اما همواره چنین نیست و موارد بسیاری وجود دارد که برای توضیح یک پدیده نیاز به بررسی عوامل کیفی داریم. معروف‌ترین مثال در این خصوص، بررسی عوامل تعیین‌کننده دستمزد است. بسیاری از این عوامل، کیفی هستند مانند تحصیلات و آموزش، جنسیت، مهارت، نوع صنعت، نژاد و ... . علاوه‌براین، موارد دیگری نیز در خصوص تغییرات کیفی وجود دارد که نقش مهمی در تحلیل رگرسیون دارند، مانند: وضع یک مالیات جدید، تغییر تعرفه‌های وارداتی و صادراتی، وضع یک قانون جدید، وقوع یک حادثه، رسیدن یک خبر خوب یا بد به بازار سهام، تغییرات فصلی، تغییرات سیاسی و غیره. اینها و موارد مشابه، موجب توجه به متغیرهای مجازی به‌عنوان شیوهٔ مناسبی برای لحاظ کردن اثر متغیرهای کیفی و تغییرات کیفی شده است. در ادامه به بررسی ماهیت متغیر مجازی و کاربردهای آن می‌پردازیم.

۲-۵ متغیر مجازی[1]

متغیر مجازی یک وضعیت کیفی یا یک تغییر کیفی را توصیف می‌کند که قابل اندازه‌گیری به‌صورت کمّی نیست. در ساده‌ترین حالت، می‌توان یک تغییر کیفی را به دو حالت تقسیم نمود:

---

[1] dummy variable

۱- تغییر کیفی رخ داده است،

۲- تغییر کیفی رخ نداده است.

و یا یک متغیر کیفی می‌تواند دو حالت داشته باشد:

۱- حالت موردنظر،

۲- سایر حالت‌ها.

در چنین مواردی، یک متغیر مجازی برای توصیف این دو حالت تعریف می‌شود. برهمین اساس متغیر مجازی طوری تعریف می‌شود که فقط دو مقدار ۰ و ۱ را اختیار می‌کند: ۱ برای حالت موردنظر و ۰ برای سایر حالت‌ها.

مثال ۱-۵: فرض کنید عوامل تعیین‌کنندهٔ دستمزد را بررسی می‌کنیم. به نظر می‌رسد که جنسیت یکی از این عوامل است. لذا نمونه موردنظر را به دو گروه زن و مرد تقسیم می‌کنیم. متغیر مجازی D را برای بررسی اثر جنسیت بر دستمزد تعریف می‌کنیم به‌گونه‌ای که $D=0$ برای زنان و $D=1$ برای مردان. لذا دستمزد را تابعی از متغیر مجازی D در نظر می‌گیریم.

مثال ۲-۵: تصور کنید که در پانزدهم آذر ماه خبری در خصوص شرکت A رسیده است که به نظر می‌رسد بازدهی آن را تحت تأثیر قرار داده است. برای بررسی تأثیر این خبر، بازدهی‌های روزانه شرکت A در ۱۰ روز قبل از خبر و ۱۰ روز بعد از خبر را گردآوری کرده‌ایم. سپس یک متغیر مجازی برای آن تعریف می‌کنیم که $D=0$ برای روزهای قبل از خبر و $D=1$ برای روزهای بعد از خبر می‌باشد.

۳-۵ رگرسیون روی متغیر مجازی

فرض کنید Y_i متغیر وابسته است که تحت تأثیر یک عامل کیفی قرار دارد. برای بررسی تأثیر این عامل، متغیر مجازی D_i را تعریف می‌کنیم. هرگاه این عامل، وجود داشته باشد $D_i=1$ و در غیر این صورت $D_i=0$ می‌باشد. رگرسیون Y_i روی D_i عبارت است از:

$$Y_i=\alpha+\beta D_i+u_i \qquad i=1,\cdots,n \tag{۵-۱}$$

$$D_i=\begin{cases}1 & \text{عامل مورد نظر وجود دارد}\\ 0 & \text{عامل مورد نظر وجود ندارد}\end{cases}$$

معادله (۱-۵) را با روش OLS برآورد می‌کنیم که بدین منظور لازم است مجموع مجذور خطاها حداقل شود. از این معادله می‌توان برای تحلیل تأثیر عامل موردنظر استفاده نمود. از آنجا که D_i برابر ۰ و ۱ است، لذا فقط دو مقدار برای Y_i به دست می‌آید.

$$D_i = 0 \Rightarrow Y_i = \alpha \quad \text{عدم وجود عامل موردنظر} \qquad (۵-۲)$$

$$D_i = 1 \Rightarrow Y_i = \alpha + \beta \quad \text{وجود عامل موردنظر}$$

بنابراین تأثیر عامل موردنظر برابر با β است.

$$\frac{\Delta Y_i}{\Delta D_i} = \beta \Rightarrow \Delta Y_i = \beta \;\; ; \;\; \Delta D_i = 1 - 0 = 1 \qquad (۵-۳)$$

اگر تأثیر عامل موردنظر را بر حسب درصد بیان کنیم، آنگاه درصد افزایش Y عبارت است از:

$$Y \text{ درصد افزایش در} = \frac{\Delta Y_i}{Y_i} = \frac{\beta}{\alpha} \qquad (۵-۴)$$

بنابراین، $\frac{\beta}{\alpha}$ بیانگر درصد افزایش در Y است که ناشی از وجود یا وقوع عامل موردنظر می‌باشد.

مثال ۳-۵: تصور کنید می‌خواهیم اثر جنسیت بر دستمزد را بررسی کنیم. بدین منظور متغیر مجازی D_i را تعریف می‌کنیم که $D = 0$ برای زنان و $D = 1$ برای مردان می‌باشد.

$$Y_i = \alpha + \beta D_i + u_i$$

نتایج این مدل عبارت است از:

$$D_i = 0 \Rightarrow Y_i = \alpha \quad \text{متوسط دستمزد زنان}$$

$$D_i = 1 \Rightarrow Y_i = \alpha + \beta \quad \text{متوسط دستمزد مردان}$$

اگر $\beta > 0$ و معنادار باشد، آنگاه می‌توان گفت که دستمزد مردان به‌طور متوسط بیشتر از دستمزد زنان است. بنابراین، افزایش دستمزد به خاطر جنسیت برابر با $\frac{\Delta Y_i}{\Delta D_i} = \Delta Y_i = \beta$ و درصد افزایش دستمزد که ناشی از جنسیت است برابر با $\frac{\Delta Y_i}{Y_i} = \frac{\beta}{\alpha}$ می‌باشد.

۵-٤ دام متغیرهای مجازی[1]

در استفاده از متغیرهای مجازی بایستی دقت شود که اگر جامعه را به دو گروه تقسیم می‌کنیم، فقط برای یک گروه می‌توان متغیر مجازی تعریف نمود. بدین منظور تصور کنید که برای گروه اول D_{1i} و برای گروه دوم D_{2i} را معرفی کنیم. از طرف دیگر می‌دانیم که عرض از مبدأ (α) ضریب متغیری مانند X_i است که تمامی مقادیر آن برابر ۱ می‌باشد.

$$Y_i = \alpha X_i + \beta D_{1i} + \gamma D_{2i} + u_i \quad , \quad X_i = 1 \tag{۵-۵}$$

متغیرهای مجازی نیز به‌صورت زیر تعریف شده‌اند:

$$D_{i2} = \begin{cases} 0 & \text{برای گروه ۱} \\ 1 & \text{برای گروه ۲} \end{cases}$$

$$D_{i1} = \begin{cases} 1 & \text{برای گروه ۱} \\ 0 & \text{برای گروه ۲} \end{cases}$$

از آنجا که $D_{i1} + D_{i2} = 1$ است لذا برابر با X_i خواهد بود که موجب همخطی کامل می‌شود و امکان برآورد ضرایب وجود نخواهد داشت. همچنین این موضوع را می‌توان به‌صورت دیگری بررسی نمود. چون $D_{i2} = 1 - D_{i1}$ است، با جایگذاری در معادله رگرسیون خواهیم داشت:

$$\begin{aligned} Y_i &= \alpha X_i + \beta D_{1i} + \gamma(1 - D_{i1}) + u_i \\ &= \alpha X_i + (\beta - \gamma) D_{i1} + \gamma + u_i \end{aligned} \tag{۵-۶}$$

γ نیز ضریب متغیری مانند Z_i است که مقادیر آن برابر ۱ است:

$$Y_i = \alpha X_i + (\beta - \gamma) D_{1i} + \gamma Z_i + u_i \tag{۵-۷}$$

چون $X_i = Z_i = 1$ است، لذا بین X_i و Z_i همخطی کامل برقرار است و در نتیجه، امکان برآورد این معادله وجود ندارد. نتیجه مهم این بحث آن است که وقتی جامعه به دو گروه تقسیم می‌شود، بایستی فقط یک متغیر مجازی تعریف گردد. در حالت کلی، اگر جامعه به m گروه تقسیم شود، بایستی $m-1$ متغیر مجازی معرفی گردد تا مشکل هم‌خطی به وجود نیاید. این موضوع معروف به دام متغیر مجازی است. البته توجه شود که در رگرسیون فاقد عرض از مبدأ، می توان m متغیر مجازی تعریف نمود.

---

[1] dummy variable trap

۴-۵ رگرسیون روی متغیر مجازی همراه با متغیر توضیحی

فرض کنید که Y_i علاوه بر X_i تابعی از یک عامل کیفی است که آن را با متغیر مجـازی D_i نشان می‌دهیم:

$$Y_i = \alpha + \beta X_i + \gamma D_i + u_i \qquad (۵-۸)$$

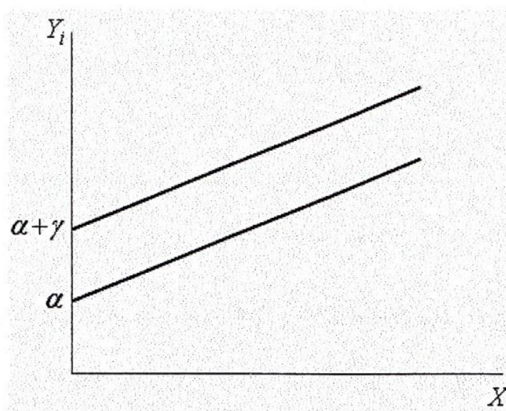
$$D_i = \begin{cases} ۱ & \text{عامل موردنظر وجود دارد} \\ ۰ & \text{عامل موردنظر وجود ندارد} \end{cases}$$

نتایج حاصل از معادله فوق به‌صورت زیر تحلیل می‌شود:

$$D_i = ۰ \Rightarrow Y_i = \alpha + \beta X_i \qquad \text{عامل مورد نظر وجود ندارد}$$

$$D_i = ۱ \Rightarrow Y_i = \alpha + \beta X_i + \gamma = (\alpha + \gamma) + \beta X_i \qquad \text{عامل موردنظر وجود دارد}$$

ملاحظه می‌شود که عامل موردنظر می‌تواند موجب انتقال خط رگرسیون شود. به‌عبارت دیگر عرض از مبدأ معادله را تغییر می‌دهد. با فرض اینکه $\gamma > ۰$ باشد، نمودار زیر را خواهیم داشت:



نمودار ۱-۵: تغییر در عرض از مبدأ

۵-۵ متغیر مجازی و تغییر شیب

عوامل کیفی که اثرشان با متغیرهای مجازی در نظر گرفته می‌شود، عـلاوه بـر عـرض از مبـدأ، می‌توانند شیب معادله را نیز تغییر دهد. در این صورت اگر فقط شیب و نه عـرض از مبـدأ را تغییـر دهد، می‌توان آن را به‌صورت زیر توصیف نمود:

$$Y_i = \alpha + \beta X_i + \theta D_i X_i + u_i \qquad (۵-۹)$$

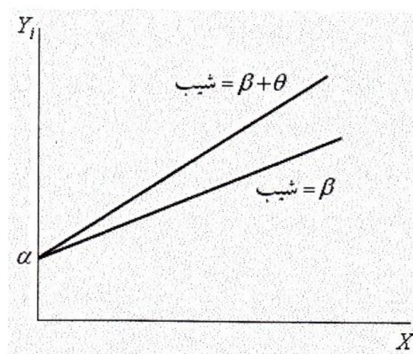
$$D_i = \begin{cases} ۱ & \text{عامل موردنظر وجود دارد} \\ ۰ & \text{عامل موردنظر وجود ندارد} \end{cases}$$

اثر عامل موردنظر بر شیب را می‌توان به‌صورت زیر بررسی نمود:

$$D_i = ۰ \Rightarrow Y_i = \alpha + \beta X_i + u_i \qquad (۵-۱۰)$$

$$D_i = ۱ \Rightarrow Y_i = \alpha + (\beta + \theta) X_i + u_i$$

بنابراین وقتی عامل یا حادثه موردنظر رخ داده است، شیب معادله رگرسیون را تغییر می‌دهد که برابر با $\beta + \theta$ می‌شود با فرض $\theta > ۰$ نمودار زیر را داریم:



نمودار ۲-۵: تغییر شیب

اگر هم شیب و هم عرض از مبدأ تغییر کند می‌توان معادله زیر را در نظر گرفت:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \theta D_i X_i + u_i \qquad (۵-۱۱)$$

$$D_i = \begin{cases} ۱ & \text{برای حالت مورد نظر} \\ ۰ & \text{برای سایر حالت‌ها} \end{cases}$$

تحلیل نتایج معادله فوق به‌صورت زیر است:

$$D_i = ۰ \Rightarrow Y_i = \alpha + \beta X_i + u_i \qquad \text{عامل یا حادثه موردنظر رخ نداده است}$$

$$D_i = ۱ \Rightarrow Y_i = (\alpha + \gamma) + (\beta + \theta) X_i + u_i \qquad \text{عامل یا حادثه موردنظر رخ داده است}$$

٦-۵ وجود دو عامل کیفی

فرض کنید که Y_i دستمزد را نشان می‌دهد که متأثر از دو عامل سواد و سابقه کارگران است. برای هر عامل، یک متغیر مجازی تعریف را به‌صورت زیر معرفی می‌کنیم:

$$Y_i = \alpha + \beta D_{1i} + \gamma D_{2i} + u_i \qquad (5\text{-}12)$$

$$D_{1i} = \begin{cases} 1 & \text{با سابقه} \\ 0 & \text{کم سابقه} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{با تحصیلات بالا} \\ 0 & \text{با تحصیلات پایین} \end{cases}$$

چهار حالت وجود دارد که عبارتند از:

دستمزد کارگر کم‌سابقه و با تحصیلات پایین $D_{1i} = D_{2i} = 0 \Rightarrow Y_i = \alpha$

دستمزد کارگر باسابقه و با تحصیلات پایین $D_{1i} = 1 , D_{2i} = 0 \Rightarrow Y_i = \alpha + \beta$

دستمزد کارگر کم‌سابقه و با تحصیلات بالا $D_{1i} = 0 , D_{2i} = 1 \Rightarrow Y_i = \alpha + \gamma$

دستمزد کارگران باسابقه و با تحصیلات بالا $D_{1i} = D_{2i} = 1 \Rightarrow Y_i = \alpha + \gamma + \beta$

با تخمین این معادله، اگر β و γ از نظر آماری معنادار باشند، نشان می‌دهد که سابقه و تحصیلات بر دستمزد تأثیر مثبت دارند. α دستمزد پایه را در حالتی نشان می دهد که $D_1 = D_2 = 0$ باشد، یعنی دستمزد فردی کم‌سابقه و با تحصیلات پایین است. β تأثیر سابقه را نشان می‌دهد که به دو صورت قابل تحلیل است.

۱- تأثیر سابقه برای «فرد با تحصیلات پایین» عبارت است از:

$$\begin{cases} D_{i1} = 0 \Rightarrow Y_i = \alpha \\ D_{i1} = 1 \Rightarrow Y_i = \alpha + \beta \end{cases}$$

$$\Delta D_{i1} = 1 \Rightarrow \Delta Y_i = \beta \;\Rightarrow\; \frac{\Delta Y_i}{Y_i} = \frac{\beta}{\alpha} \qquad (5\text{-}13)$$

۲- تأثیر سابقه برای «فرد با تحصیلات بالا» عبارت است از:

$$\begin{cases} D_{i1} = 0 \Rightarrow Y_i = \alpha + \gamma \\ D_{i1} = 1 \Rightarrow Y_i = \alpha + \gamma + \beta \end{cases}$$

$$\Delta D_{i1} = 1 \Rightarrow \Delta Y_i = \beta \Rightarrow \frac{\Delta Y_i}{Y_i} = \frac{\beta}{\alpha + \gamma} \qquad (5\text{-}14)$$

۷-۵ تأثیر متقابل دو عامل کیفی

نکته دیگری که می‌توان به بحث قبلی اضافه نمود آن است که اگر دو عامل کیفی به‌طور همزمان وجود داشته باشند، اثر متقابل آنها را بایستی به‌گونه دیگری تحلیل نمود. به‌عنوان مثال سابقه و تحصیلات ممکن است اثر مضاعفی بر افزایش دستمزد داشته باشند. بدین منظور، مدل زیر را برای لحاظ کردن اثرات متقابل ارائه می‌شود:

$$Y_i = \alpha + \beta D_{1i} + \gamma D_{2i} + \theta D_{1i} D_{2i} + u_i \qquad (۵-۱۵)$$

حاصل‌ضرب $D_{1i}D_{2i}$ بیانگر اثر متقابل دو عامل کیفی است: سابقه و تحصیلات. در اینجا چهار حالت وجود دارد:

۱) $D_{1i} = D_{2i} = 0 \Rightarrow Y_i = \alpha$ متوسط دستمزد کارگر کم‌سابقه و با تحصیلات پایین

۲) $D_{1i} = 1 , D_{2i} = 0 \Rightarrow Y_i = \alpha + \beta$ متوسط دستمزد کارگر باسابقه و با تحصیلات پایین

۳) $D_{1i} = 0 \Rightarrow D_{2i} = 1 \Rightarrow Y_i = \alpha + \gamma$ متوسط دستمزد کارگر کم‌سابقه و با تحصیلات بالا

۴) $D_{1i} = D_{2i} = 1 \Rightarrow Y_i = \alpha + \beta + \gamma + \theta$ متوسط دستمزد کارگر باسابقه و با تحصیلات بالا

θ اثر متقابل سابقه و تحصیلات بر دستمزد را نشان می‌دهد. در حالی که β اثر سابقه و γ اثر تحصیلات را به‌طور جداگانه نشان می‌دهند.

۸-۵ متغیرهای مجازی و تغییر ساختار

فرض کنید که داده‌های گردآوری شده مربوط به دوره ۱۳۸۰-۱۳۴۰ باشد. در طول این دوره وقایع زیادی رخ داده است، مانند وقوع انقلاب اسلامی در سال ۱۳۵۷. ممکن است تصور بر این باشد که وقوع چنین وقایعی می‌تواند رگرسیون را دچار تغییر کند. به‌عنوان مثال آن را منتقل کند و یا شیب آن را تغییر دهد. با استفاده از متغیرهای مجازی می‌توان اثر این وقایع را در نظر گرفت. بدین منظور فرض کنید که معادله رگرسیون به‌صورت $Y_t = \alpha + \beta X_t + u_t$ است و می‌خواهیم اثر واقعه موردنظر را (مثلاً در سال ۱۳۵۷) به این معادله اضافه کنیم.

$$Y_t = \alpha + \beta X_t + \gamma D_t + \theta D_t X_t + u_t \qquad (۵-۱۶)$$

$$D_t = \begin{cases} 0 & t < 1357 \\ 1 & t \geq 1357 \end{cases}$$

بنابراین، دو معادله رگرسیون داریم، یکی برای قبل از سال ۱۳۵۷ و دیگری برای بعد از آن:

$$D_t = \circ \Rightarrow Y_t = \alpha + \beta X_t + u_t \quad , \quad t < ۱۳۵۷ \qquad (۵-۱۷)$$

$$D_t = ۱ \Rightarrow Y_t = (\alpha + \gamma) + (\beta + \theta) X_t + u_t \quad , \quad t \geq ۱۳۵۷$$

حالت دیگر آن است که اثر واقعه موردنظر را فقط در زمان وقوع آن، لحاظ کنیم. در این صورت معادله زیر را تعریف می‌کنیم:

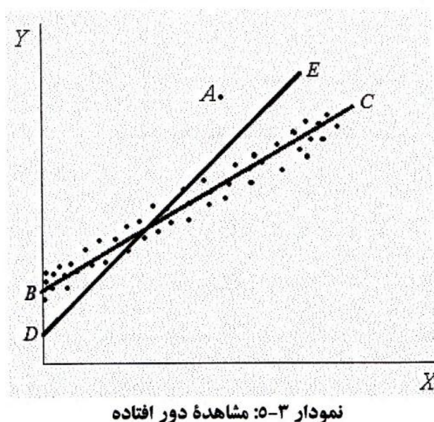
$$Y_t = \alpha + \beta X_t + \gamma D_t + u_t$$

$$D_t = \begin{cases} ۱ & D_t = ۱۳۵۷ \\ \circ & D_t \neq ۱۳۵۷ \end{cases} \qquad (۵-۱۸)$$

بدین ترتیب اثر وقوع انقلاب در سال ۱۳۵۷ توسط γ نشان داده می‌شود ولی در سالهای قبل و بعد از ۱۳۵۷ توسط معادله $Y_t = \alpha + \beta X_t + u_t$ نشان داده می‌شود. γ اثر واقعه موردنظر است که به‌صورت تغییر در یک سال معین می‌باشد.

۵-۹ متغیرهای مجازی و مشاهدات دور افتاده[1]

مشاهدات دور افتاده به دلایل مختلفی وجود دارند. یک مشاهده دور افتاده به هر دلیلی که باشد، خارج از محدودهٔ عادی مشاهدات قرار دارد.



نمودار ۳-۵: مشاهدهٔ دور افتاده

[1] outlier

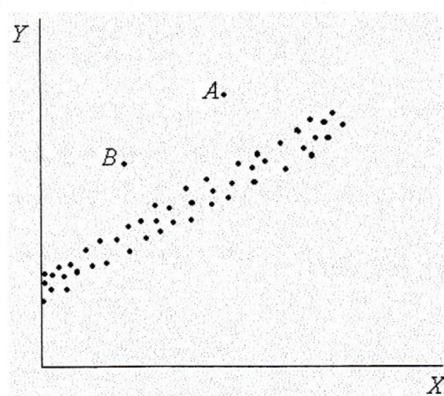
نقطه A یک مشاهدهٔ دورافتاده را نشان می‌دهد که فاصلهٔ غیرعادی با بقیه مشاهدات دارد. این مشاهده ممکن است بیانگر وقوع یک حادثه بوده که نمی‌توان آن را نادیده گرفت. وجود این مشاهده، خط رگرسیون را از BC به DE تغییر می‌دهد. این موجب افزایش خطا و کاهش R^2 می‌شود. از طرف دیگر ممکن است در نظر گرفتن اثر حادثه‌ای که فقط در نقطه A مشاهده می‌شود، برای ما مهم باشد. در این صورت می‌توان برای نقطه A یک متغیر مجازی تعریف شود:

$$Y_i = \alpha + \beta X_i + \gamma D_i + u_i \qquad (۵-۱۹)$$

$$D_i = \begin{cases} ۱ & \text{برای نقطه A} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

بنابراین، γ نشان‌دهنده اثر واقعه‌ای است که نقطه A را به‌وجود آورده است.

حال اگر مشاهدات دور افتاده به‌صورت نمودار زیر باشد، می‌توان اثر آن را با معرفی یک متغیر مجازی در نظر گرفت.



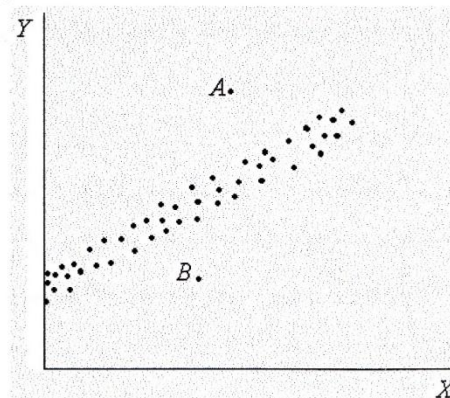
نمودار ٤-۵: مشاهدات دور افتادهٔ همسو

$$Y_i = \alpha + \beta X_i + \gamma D_i + u_i \qquad (۵-۲۰)$$

$$D_i = \begin{cases} ۱ & \text{برای نقطه A و B} \\ ۰ & \text{برای سایر نقاط} \end{cases}$$

در اینجا چون دو نقطه A و B در یک راستا قرار دارند (هر دو بالای مشاهدات اصلی، قرار دارند) معرفی یک متغیر مجازی کفایت می‌کند و اثر هر دو نقطه را در نظر می‌گیرد.

اما اگر مشاهدات دور افتاده به‌صورت زیر باشند، وضعیت متفاوتی خواهد داشت.



نمودار ۵-۵: مشاهدات دور افتادهٔ ناهمسو

در این حالت اگر یک متغیر مجازی برای نقاط A و B معرفی کنیم. ضریب آن (γ) بی‌معنی خواهد شد و نمی‌تواند اثر آنها را در نظر بگیرید. زیرا نقطه A اثر مثبت و نقطه B اثر منفی داشته است. به عبارت دیگر، نقطه A رگرسیون را به بالا و نقطه B رگرسیون را به پایین انتقال می‌دهد. بدیهی است که این دو، همدیگر را خنثی می‌کنند و ضریب متغیر مجازی را بی‌معنی خواهد کرد. بنابراین لازم است برای هر یک از این دو نقطه، متغیرهای مجازی جداگانه‌ای را تعریف کنیم. به عبارت دیگر برای تغییرات مثبت (شوک‌های مثبت) یک متغیر مجازی و برای تغییرات منفی (شوک‌های منفی) نیز یک متغیر مجازی تعریف نماییم.

$$Y_i = \alpha + \beta X_i + \gamma D_{1i} + \theta D_{2i} + u_i \qquad (۵-۲۱)$$

$$D_{1i} = \begin{cases} 1 & \text{نقطه A} \\ 0 & \text{سایر نقاط} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{نقطه B} \\ 0 & \text{سایر نقاط} \end{cases}$$

بدین ترتیب اثر مشاهدهٔ نقطه A که ناشی از یک عامل معین بوده توسط γ و اثر مشاهدهٔ نقطه B نیز که ناشی از عامل معین دیگری است توسط θ نشان داده می‌شود. به‌عنوان مثال نقاطی مانند A می‌تواند رسیدن خبرهای خوب به بازار سهام و نقاطی مانند B می‌تواند رسیدن خبرهای بد را نشان دهند. این بحث را می‌توان در موارد زیادی به‌کار برد. به‌عنوان مثال تأثیر شوک‌های مثبت و

منفی که به یک بازار وارد می‌شود یا تأثیر شوک‌های افزایش و کاهش قیمت نفت بر رشد اقتصادی.

۱۰-۵ متغیرهای مجازی و تغییرات فصلی

تغییرات فصلی نیز یکی از عوامل کیفی است که اثر آن به‌ویژه بر تقاضای کالاهای مختلف یا بر تولید کالاها می‌تواند درخور توجه باشد. فرض کنید که معادله رگرسیون به‌صورت $Y_i = \alpha + \beta X_i + u_i$ است. اثر تغییرات فصلی را با استفاده از متغیرهای مجازی می‌توان وارد این معادله کرد:

$$Y_i = \alpha + \beta X_i + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \gamma_4 D_{4i} + u_i \qquad (۵-۲۲)$$

$$D_2 = \begin{cases} 1 & \text{فصل تابستان} \\ 0 & \text{سایر فصل‌ها} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{فصل پاییز} \\ 0 & \text{سایر فصل‌ها} \end{cases}$$

$$D_4 = \begin{cases} 1 & \text{فصل زمستان} \\ 0 & \text{سایر فصل‌ها} \end{cases}$$

توجه شود که در اینجا چهار فصل داریم ولی سه متغیر مجازی تعریف کرده‌ایم. در واقع فصل بهار را به‌عنوان مبنا در نظر گرفته و برای آن متغیر مجازی معرفی نکرده‌ایم. اگر برای فصل بهار نیز متغیر مجازی D_1 را تعریف می‌کردیم، آنگاه حاصل جمع چهار متغیر مجازی در هر شرایطی برابر ۱ می‌شد که با ضریب ثابت α هم‌خطی کامل پیدا می‌کرد و دچار دام متغیر مجازی می‌شدیم.

براساس معادله (۲۲-۵) برای هر یک از فصل‌ها، معادله رگرسیون عبارت است از:

۱) فصل بهار $D_1 = D_2 = D_3 = 0 \Rightarrow Y_i = \alpha + \beta X_i + u_i$

۲) فصل تابستان $D_2 = 1, D_3 = D_4 = 0 \Rightarrow Y_i = (\alpha + \gamma_2) + \beta X_i + u_i$

۳) فصل پاییز $D_2 = 0, D_3 = 1, D_4 = 0 \Rightarrow Y_i = (\alpha + \gamma_3) + \beta X_i + u_i$

۴) فصل زمستان $D_2 = D_3 = 0, D_4 = 1 \Rightarrow Y_i = (\alpha + \gamma_4) + \beta X_i + u_i$

نتایج فوق نشان می‌دهد که تفاوت فصل‌ها توسط عرض از مبدأها مشخص شده است که موجب انتقال منحنی به سمت بالا یا پایین می‌شود.

۱۱-۵ توابع لگاریتمی و متغیر مجازی

توابعی که تاکنون معرفی کردیم، خطی بودند. اگر متغیر وابسته به‌صورت لگاریتمی باشد، آنگاه تحلیل نتایج نیز متفاوت خواهد بود. بدین منظور معادله زیر را در نظر بگیرید:

$$\ln Y_i = \alpha + \beta D_i + u_i \qquad (۵-۲۳)$$

اگر $D_i = ۰$ باشد، $\ln Y_i = \alpha$ و اگر $D_i = ۱$ باشد $\ln Y_i = \alpha + \beta$ است. بنابراین تغییر در لگاریتم Y_i برابر با β است:

$$\frac{\Delta \ln Y_i}{\Delta D_i} = \Delta \ln Y_i = \beta \qquad (۵-۲۴)$$

از طرف دیگر، تغییر Y_i برابر است با:

$$D_i = ۰ \Rightarrow \ln Y_i = \alpha \Rightarrow Y_i = e^{\alpha}$$

$$D_i = ۱ \Rightarrow \ln Y_i = \alpha + \beta \Rightarrow Y_i = e^{\alpha+\beta} = e^{\alpha} e^{\beta}$$

$$\Delta Y_i = e^{\alpha} e^{\beta} - e^{\alpha} = e^{\alpha}(e^{\beta} - ۱)$$

و درصد تغییر در Y برابر است با:

$$\frac{\Delta Y_i}{Y_i} = \frac{e^{\alpha}(e^{\beta} - ۱)}{e^{\alpha}} = e^{\beta} - ۱ \qquad (۵-۲۵)$$

از طرف دیگر می‌دانیم که در حالت پیوسته، $d \ln Y_i = \frac{dY_i}{Y_i}$ است و لذا با مشتق‌گیری از (۵-۲۳) خواهیم داشت:

$$d \ln Y_i = \beta \Rightarrow \frac{dY_i}{Y_i} = \beta \qquad (۵-۲۶)$$

تفاوت (۵-۲۵) و (۵-۲۶) در این است که در محاسبه (۵-۲۶) فرض کرده‌ایم متغیرها پیوسته و قابل مشتق‌گیری هستند. لذا فرمول (۵-۲۶) به‌صورت تقریبی است و نه دقیق. به‌هرحال

این دو، تقریباً برابرند و با استفاده از بسط تیلور می‌توان نشان داد که $\beta \cong e^{\beta} - 1$ است.[1]

۵-۱۲ قیمت‌گذاری عوامل کیفی

در برخی از کاربردهای اقتصادی و مالی می‌توان از متغیرهای مجازی به‌عنوان روشی برای قیمت‌گذاری عوامل کیفی استفاده نمود. به‌عنوان مثال در معادله تعیین دستمزد، اثر دستمزد را تابعی از سابقه بدانیم، رابطه زیر را برای آن می‌نویسیم:

$$Y_i = \alpha + \beta D_i + u_i \qquad (۵-۲۷)$$

$$D_i = \begin{cases} 1 & \text{با سابقه} \\ 0 & \text{بدون سابقه} \end{cases}$$

اگر $D_i = 0$ باشد آنگاه دستمزد برابر با $Y_i = \alpha$ است. حال اگر $D_i = 1$ باشد آنگاه دستمزد برابر با $Y_i = \alpha + \beta$ می‌باشد. بنابراین افزایش دستمزد بابت سابقه برابر با β است. β را می‌توان قیمت سابقه دانست. زیرا مقدار دستمزدی است که به‌خاطر داشتن سابقه پرداخت می‌شود. به‌عنوان مثال دیگر، در مدل‌های تقاضا، قیمت تابعی از عوامل و متغیرهای مختلفی است که بعضی از آنها کیفی هستند. در این مدل‌ها، قیمت معادل با تمایل به پرداخت مصرف‌کننده بابت این عوامل است. به‌عنوان مثال مدل زیر را در نظر بگیرید:

$$P_i = \alpha + \beta X_i + \gamma D_i + u_i \qquad (۵-۲۸)$$

معادله فوق یک معادله تقاضا است که بر حسب قیمت، تنظیم شده است. قیمتی که متقاضی می‌پردازد یا به‌طور کلی، قیمت بازار تابعی از خصوصیات کالا و عوامل تعیین‌کننده آن است. در اینجا فقط دو عامل را ذکر کرده‌ایم که یکی متغیر کمّی X_i و دیگری متغیر یا عامل کیفی است که D_i آن را نشان می‌دهد. اگر D_i از ۰ به ۱ افزایش یابد، قیمت در هر سطحی از X_i افزایش می‌یابد که مقدار افزایش آن برابر با γ است. لذا γ معادل با قیمت یا ارزش عامل موردنظر از

---

[1] بسط تیلور یا تقریب مرتبه اول برای تابع $f(x)$ حول $x = 0$ برابر با $f(x) \cong f(0) + f'(0)x$ می‌باشد. مشابه این رابطه برای تابع $f(\beta) = e^{\beta} - 1$ حول $\beta = 0$ برابر است با:

$$f(\beta) \cong f(0) + f'(0)\beta = (e^0 - 1) + e^0 \beta = \beta$$

دیدگاه مصرف‌کننده است. این نوع از معادلات در رابطه با تقاضای مسکن کاربرد زیادی دارد. طبق این معادله، قیمت مسکن تابعی از عوامل کیفی مانند، دسترسی به بازار، داشتن سیستم گرمایش و سرمایش، موقعیت مکانی، شلوغی و غیره. هر یک از این عوامل نقشی در تعیین قیمت مسکن دارند که ضریب هر یک از آنها، معادل با ارزش (مثبت یا منفی) عامل موردنظر است.

متغیرهای مجازی در Eviews فایل data8

فرض کنید مدل زیر را می‌خواهیم برآورد کنیم:

$$Y_i = \alpha_1 + \beta_1 X_i + \alpha_2 D_{1i} + \beta_2 D_{1i} X_i + u_i$$

$$D_{1i} = \begin{cases} 1 & X < 100 \\ 0 & X \geq 100 \end{cases}$$

برای تخمین این مدل می‌توان متغیر مجازی D_{1i} را با دستور data D1 تعریف نمود. اما راه ساده‌تر آن است که به‌صورت زیر عمل کنیم.

برای برآورد این معادله، مراحل زیر را انجام می‌دهیم:

۱- ابتدا متغیر D2 را با دستور genr D2=1 ایجاد می‌کنیم.

۲- معادله مورد نظر را به صورت زیر برآورد می‌کنیم:

```
LS Y C X D2*X<100 (D2*X<100)*X
```

در واقع ضریب D2*X<100 معادل با ضریب D_{1i} (یعنی α_2) و ضریب (D2*X<100)*X معادل با ضریب $D_{1i}X_i$ (یعنی β_2) است.

Equation: UNTITLED Workfile: DATA20::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 12/26/13 Time: 11:05
Sample: 1351 1360
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-13.24402	2.000893	-6.619054	0.0006
X	0.194258	0.014948	12.99525	0.0000
D2*X<100	16.57112	2.236107	7.410701	0.0003
(D2*X<100)*X	-0.167156	0.022681	-7.369803	0.0003

R-squared	0.983085	Mean dependent var	9.300000
Adjusted R-squared	0.974628	S.D. dependent var	5.538752
S.E. of regression	0.882252	Akaike info criterion	2.876497
Sum squared resid	4.670214	Schwarz criterion	2.997531
Log likelihood	-10.38248	Hannan-Quinn criter.	2.743723
F-statistic	116.2387	Durbin-Watson stat	2.391669
Prob(F-statistic)	0.000011		

بنابراین معادله تخمینی عبارت است از:

$$\hat{Y}_i = -13/24 + 0/19 X_i + 16/57 D_{1i} - 0/17 D_{2i} X_i$$

مسائل

۱- به نظر می‌رسد که مقدار فروش یک بنگاه در فصول مختلف دارای تفاوت معناداری است. برای فروش این بنگاه یک معادله رگرسیون با استفاده از متغیرهای مجازی به‌منظور لحاظ کردن اثرات فصلی معرفی کرده و موارد زیر را پاسخ دهید.

الف) آزمون یکسان بودن فروش در فصل‌های مختلف.

ب) آزمون برابری فروش در فصل بهار و پاییز.

ج) آزمون برابری فروش در فصل بهار و پاییز و برابری فروش در تابستان و پاییز.

۲- تقاضای بنزین تابعی از قیمت آن است. لذا دولت می‌خواهد با افزایش قیمت، مصرف بنزین را کاهش دهد. اما استدلال می‌شود که چون قیمت بنزین پایین است، نمی‌توان از سیاست‌های قیمتی برای کنترل مصرف استفاده نمود؛ مگر آنکه قیمت بنزین از یک مقدار آستانه (P^*) عبور کند. این وضعیت را روی نمودار نشان داده و یک معادله تقاضای بنزین را برای آن معرفی کنید.

۳- نظریه مصرف دوزنبری بیان می‌کند که رفتار مصرفی در واکنش به تغییرات درآمد، قرینه نمی‌باشد. بدین معنی که به افزایش‌های درآمد، واکنش بیشتری نشان می‌دهد ولی به کاهش‌های درآمد، واکنش کمتری نشان می‌دهد. این رفتار مصرفی را با استفاده ازمتغیرهای مجازی توصیف کنید.

۴- معادله مصرف در مسئله ۳ را چگونه با Eviews برآورد می‌کنید.

۵- مشاهدات مربوط به Y_t در دوره ۸۰-۱۳۵۰ به‌صورت شش ماهه در د سترس است. داده‌های شش‌ماهه اول متفاوت از شش ماهه دوم است. با استفاده از متغیرهای مجازی اثر تفاوت فصلی را بر معادله $Y_t=\alpha+\beta X_t+u_t$ مدل‌سازی کنید.

۶- در معادله $Y_i=\alpha F_i+\beta M_i+u_i$ درآمد روی جنسیت برازش شده است. F_i برای زنان و M_i برای مردان برابر با ۱ است. ثابت کنید که $\hat{\alpha}$ برابر با میانگین درآمد زنان و $\hat{\beta}$ برابر با میانگین درآمد مردان است.

۷- معادله $Y_i=\alpha+\beta D_i+u_i$ بیانگر رگرسیون درآمد روی جنسیت می‌باشد. D_i متغیر مجازی است که برای مردان برابر با ۱ و برای زنان برابر با ۰ است. متوسط درآمد زنان و مردان چقدر است.

۸- فرض کنید که داده‌های تقاضا کالای Q و قیمت آن (P) و همچنین درآمد مصرف‌کنندگان (I) به‌صورت زیر باشد. توجه شود که مقدار تقاضا به صورت فصلی ولی قیمت و درآمد به‌صورت سالانه در دسترس بوده است.

سال	فصل	Q	P	I
۱۳۸۰		۱۰۰	۵۰	۱۰۰۰
		۱۵۰		
		۱۰۰		
		۱۲۰		
۱۳۸۱		۱۱۰	۶۰	۱۲۰۰
		۱۵۰		
		۱۲۰		
		۱۳۰		
۱۳۸۲		۱۱۵	۶۵	۱۵۰۰
		۱۵۰		
		۱۱۰		
		۱۲۰		
۱۳۸۳		۱۲۰	۸۰	۱۸۰۰
		۱۶۰		
		۱۰۰		
		۱۴۰		
۱۳۸۴		۱۳۰	۹۰	۲۰۰۰
		۱۷۰		
		۱۲۰		
		۱۳۰		

الف) متوسط تقاضا برای هر سال را حساب کرده و معادله تقاضا را بر اساس داده‌های سالانه برآورد کنید.

ب) اثر تفاوت‌های فصلی را در معادله تقاضا آزمون کنید.

۹- معادله $Y_i = \alpha + \beta D_i + u_i$ بیانگر رگرسیون درآمد روی جنسیت است. آزمون برابری درآمد مردان و زنان را چگونه انجام می‌دهید.

۱۰- معادله $Y_i = \alpha F_i + \beta M_i + u_i$ بیانگر رگرسیون درآمد روی جنسیت است (F_i برای زنان و M_i برای مردان). آزمون برابری درآمد مردان و زنان را چگونه انجام می‌دهید.

۱۱- اگر مؤسسه‌ای به کارکنان خود در ساعات معمول، دستمزد ثابتی به‌ازای هر ساعت بپردازد، ولی بابت اضافه‌کاری آنها به ازای هر ساعت، دستمزد ثابت و بالاتری بپردازد، یک معادله رگرسیون برای توصیف این وضعیت به کمک متغیرهای مجازی معرفی کنید. همچنین اگر به‌ازای هر ساعت اضافه کاری، دستمزد بالاتر و متغیر بپردازد، یک معادله رگرسیون برای توصیف این وضعیت به کمک متغیرهای مجازی معرفی کنید.

ضمیمه فصل ۵: متغیرهای مجازی در Stata

متغیرهای مجازی در Stata

استفاده از متغیرهای مجازی مانند متغیرهای معمولی است. اما مسئله مهم این است که متغیر مجازی را برای چه موضوعی می‌سازیم و چه هدفی را دنبال می‌کنیم. بعد از مشخص نمون موضوع و هدف مورد نظر می‌توان متغیر مجازی را تعریف کرده و سپس مقادیر آن را که برابر با صفر و یک می‌باشند وارد نمود.

به عنوان مثال فرض کنید مدل زیر را می‌خواهیم برآورد کنیم:

$$Y_i = \alpha_1 + \beta_1 X_i + \alpha_2 D_{1i} + \beta_2 D_{1i} X_i + u_i$$

$$D_{1i} = \begin{cases} 1 & X < 100 \\ 0 & X \geq 100 \end{cases}$$

ابتدا متغیر مجازی D را با فرمان زیر می‌سازیم:

```
gen d=(x<100)
```

سپس حاصل‌ضرب متغیرهای D و X را با فرمان زیر حساب می‌کنیم:

```
gen  dx=d*x
```

سپس معادله را با فرمان زیر تخمین می‌زنیم:

```
reg  x d dx
```

```
. reg  y x d dx

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  3,     6) =  116.24
       Model |  271.429786     3  90.4765953           Prob > F      =  0.0000
    Residual |  4.67021419     6  .778369032           R-squared     =  0.9831
-------------+------------------------------           Adj R-squared =  0.9746
       Total |       276.1     9  30.6777778           Root MSE      =  .88225

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1942584   .0149484    13.00   0.000     .1576809    .2308358
           d |   16.57112   2.236107     7.41   0.000     11.09956    22.04268
          dx |  -.1671556   .0226811    -7.37   0.000    -.2226543   -.1116568
       _cons |  -13.24402   2.000893    -6.62   0.001    -18.14003   -8.34801
------------------------------------------------------------------------------
```

متغیرهای مجازی را به اشکال دلخواه می‌توان ایجاد کرد که برخی از آنها عبارتند از:

```
gen d=(t=1357)

gen d=(t<=1357)

gen d=(t>1357)

gen d=(x>=250)

gen d=(x=120)
```

فصل ششم

آزمون‌های خاص: تصریح مدل، تغییر ساختاری و علیّت

۱-٦ مقدمه

در تخمین معادلات، مسائل و موضوعات خاصی وجود دارد که متفاوت از مباحث قبلی است. بدین منظور در این فصل به موضوعات خاصی می‌پردازیم که معمولاً در تخمین ضرایب معادلات با آن مواجه می‌شویم. این موضوعات در اغلب موارد، موجب تورش در ضرایب برآوردی و یا بی‌اعتبار شدن آزمون‌های فرضیه می‌شود که لازم است اقداماتی در جهت رفع آنها صورت گیرد. این موضوعات شامل همخطی و مشکلات مربوط به آن، فرم تابعی غلط، حذف متغیرهای مهم، افزودن متغیرهای نامربوط، ثبات ضرایب، خطای اندازه‌گیری و علّیت می‌باشد.

۲-٦ همخطی

۱-۲-٦ مفهوم همخطی

در تخمین پارامترها با استفاده از روش OLS فرض ضمنی این است که متغیرهای توضیحی با یکدیگر همبستگی خطی ندارد. همبستگی خطی بین متغیرها موسوم به همخطی است که در فصل سوم اشاره مختصری به آن شد. همان‌طور که قبلاً دیدیم تخمین ضرایب معادله رگرسیون به‌صورت $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ می‌باشد. وضعیت همخطی بستگی به ماهیت ماتریس $\mathbf{X'X}$ و یا ماتریس $\mathbf{X}$ دارد. در این خصوص، سه حالت وجود دارد:

۱- اگر همخطی کامل وجود داشته باشد بدین معنی است که بین دو یا چند ستون از ماتریس $\mathbf{X'X}$ ترکیب خطی وجود دارد (یعنی بین دو یا چند متغیر توضیحی، رابطه خطی وجود دارد) و لذا دترمینان ماتریس $\mathbf{X'X}$ برابر با صفر شده و نمی‌توان ضرایب معادله رگرسیون را برآورد نمود. این حالت بیانگر آن است که لااقل یکی از متغیرهای توضیحی، ترکیب خطی از سایر متغیرهای توضیحی است. به‌عبارت دیگر در ماتریس $\mathbf{X}$ یکی از ستون‌ها ترکیبی از یک یا چند ستون دیگر است. به‌عنوان مثال اگر X_{2t} ترکیب خطی از سایر متغیرها باشد، بدان معنا است که X_{2t} حاوی هیچ اطلاعات مفید و جدیدی نیست، زیرا اطلاعات آن صرفاً از متغیرهای دیگر به‌دست آمده است. به‌عوان مثال برای برآورد K ضریب، بایستی K ستون (متغیر) مستقل در ماتریس $\mathbf{X}$ داشته باشیم. لذا اگر یکی از ستون‌ها (متغیرها) ترکیب خطی از بقیه ستون‌ها باشد، آنگاه یکی از اطلاعات ما، مفید نخواهد بود. باین بدان معنا است که اطلاعات و داده‌ها کمتر از تعداد ضرایب است و لذا امکان برآورد ضرایب وجود نخواهد داشت.

۲- حالت دیگر این است که هیچ‌گونه رابطه خطی بین متغیرها وجود نداشته باشد. در این حالت، کوواریانس یا ضریب همبستگی بین متغیرهای توضیحی صفر می‌شود. همان‌طور که در فصل سوم دیدیم، ماتریس $\mathbf{X'X}$ تبدیل به یک ماتریس قطری شده و تخمین هر یک از ضرایب هیچ ارتباطی با سایر ضرایب ندارد.

۳- حالت عمومی و متداول این است که همبستگی بین متغیرهای توضیحی بین صفر و یک می‌باشد. بدیهی است که در این حالت، مشکل هنگامی به‌وجود می‌آید که همخطی شدید باشد و همبستگی بین متغیرهای توضیحی نسبتاً بزرگ باشد. اگر چه همخطی مشکلاتی را در تخمین پارامترها ایجاد می‌کند، ولی بر خواص تخمین‌زننده‌های OLS تأثیری ندارد.

۶-۲-۲ مشکلات ناشی از همخطی

در فصل سوم دیدیم که در یک رگرسیون دو متغیره اگر بین دو متغیر توضیحی همخطی وجود داشته باشد ضریب همبستگی آنها (r_{23}) افزایش می‌یابد و این موجب افزایش واریانس تخمین‌زننده‌های OLS می‌شود. افزایش واریانس نیز به نوبه خود باعث می‌شود که فاصله اطمینان پارامترها، عریض‌تر شود. از طرف دیگر چون واریانس تخمین‌زننده‌ها افزایش می‌یابد، لذا موجب بی‌معنی شدن ضرایب می‌شود و t آنها را غالباً کاهش می‌دهد.

در صورت وجود همخطی، علی‌رغم اینکه مقادیر t کوچک است، ولی R^2 بزرگ می‌باشد. وجود این وضعیت بدان معنا است که یک رابطه خطی بین متغیرهای توضیحی وجود دارد که توانسته است تغییرات متغیر وابسته را به‌خوبی توضیح دهد. توجه شود که مقدار t برای معنادار بودن یک ضریب به‌کار می‌رود، ولی مقدار F برای معنی‌دار بودن همه ضرایب است. ممکن است این دو (یعنی F یا R^2 بزرگ ولی t کوچک) متناقض به‌نظر آید. ولی توجه داریم که در اینجا مسئلهٔ مهم رابطه خطی بین متغیرهای توضیحی است. بدین منظور ابتدا کوواریانس بین $\hat{\beta}_2$ و $\hat{\beta}_3$ را در نظر بگیرید که عبارت است از:

$$\operatorname{cov}(\hat{\beta}_2 , \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{i1}^2 \sum x_{i2}^2}} \qquad (۶-۱)$$

از این فرمول واضح است که با افزایش r_{23}، کوواریانس تخمین‌زنندها افزایش می‌یابد. بنابراین، در صورت همخطی شدید، کوواریانس تخمین‌زننده‌ها بزرگ می‌شود و لذا حذف یا ورود یک متغیر موجب تغییر در مقدار ضرایب می‌شود. به‌عبارت دیگر تخمین پارامترها به‌شدت به‌همدیگر وابسته هستند و لذا نمی‌توان مقدار دقیق این پارامترها را برآورد نمود.

۳-۲-۶ شناسایی همخطی

مشکلی که همخطی ایجاد می‌کند آن است که با افزایش ضریب همبستگی بین دو متغیر، واریانس تخمین‌زننده‌های آن‌ها افزایش یابد و مقدار t را کوچک می‌کند در این حالت استنباط ما این است که ضریب مورد نظر بی‌معنی است و باید از مدل حذف شود. این استنباط نادرست بوده و حذف یک متغیر موجب می‌شود تا برآورد ما از سایر ضرایب نیز دچار تورش شود. بدین ترتیب، همخطی هنگامی ایجاد مشکل می‌کند که یک پارامتر به غلط اعتبار خود را از دست بدهد. آزمون‌های تشخیص همخطی نیز باید قادر باشند تا بی‌معنی بودن ضرایب به‌دلیل همبستگی بین متغیرهای توضیحی را تشخیص دهند. از آنجا که تشخیص اینکه یک پارامتر به دلیل همخطی، بی‌معنی شده است کار دشواری است، لذا نمی‌توان به یک آزمون قطعی برای تشخیص همخطی دست یافت. ولی برخی قواعد و روش‌های توافقی وجود دارند که برای تشخیص همخطی مورد استفاده قرار می‌گیرند.

یکی از معیارهای ساده جهت شناسایی همخطی استفاده از ضرایب همبستگی بین متغیرهای توضیحی است. اگر ضرایب همبستگی بین متغیرهای توضیحی، نسبتاً بزرگ باشد بیانگر همخطی نسبتاً شدید است. اما اگر ضرایب همبستگی کوچک باشند بدین معنی نیست که همخطی وجود ندارد. به‌عبارت دیگر علیرغم پایین بودن این ضرایب، ممکن است همخطیِ نسبتاً بالایی در مدل وجود داشته باشد. در چنین شرایطی، یک پیشنهاد این است که از ضرایب همبستگی جزئی استفاده شود که در فصل دوم راجع به آنها بحث شد. معمولاً گفته می‌شود که اگر برای یک معادله رگرسیون، ضریب همبستگی ساده بین متغیرهای توضیحی بیش از $\sqrt{R^2}$ باشد، همخطی شدید است.

محاسبه ضرایب همبستگی ساده در Eviews **فایل** data2

نحوهٔ محاسبه «ضرایب همبستگی ساده» با استفاده از نرم‌افزار Eviews در فصل دوم توضیح داده شد. به‌عنوان مثال در برآورد تابع تولید، متغیرهای تولید ناخالص داخلی (Y)، نیروی کار (L) و سرمایه (K) را داریم که ضرایب همبستگی آنها عبارت است از:

	Y	K	L
Y	1	0.894944	0.917590
K	0.894944	1	0.867922
L	0.917590	0.867922	1

جدول فوق نشان می‌دهد که ضریب همبستگی بین L و K حدود ۰/۸۶۸ می‌باشد که نسبتاً بالا است و نشان از همخطی است.
از طرف دیگر اگر به‌جای تابع تولید، از معادله نرخ رشد استفاده کنیم، در این صورت معمولاً همخطی بین متغیرهای توضیحی کاهش می‌یابد. ضرایب همبستگی بین نرخ رشد متغیرها عبارتند از:

	GY	GK	GL
GY	1	0.4301294	0.555101
GK	0.4301294	1	0.243341
GL	0.5551013	0.2433416	1

جدول فوق نشان می‌دهد که ضریب همبستگی بین نرخ رشد کار و سرمایه حدود ۰/۲۴۳ می‌باشد که نشان می‌دهد در چنین حالتی همخطی قابل اغماض است.

روش دیگر برای شناسایی همخطی این است که هر یک از متغیرهای توضیحی را روی بقیه متغیرهای توضیحی برازش نموده و ضریب تعیین آن را حساب کرده و با ضریب تعیین مدل اصلی مقایسه کنیم. اگر ضریب تعیین معادله مورد نظر از ضریب تعیین مدل اصلی بیشتر شد، در این

صورت همخطی شدید و غیر قابل اغماض است. به‌عنوان مثال فرض کنید که مدل زیر را داشته باشیم:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_K X_{Kt} + u_t = \beta_1 + \sum_{i=2}^{K} \beta_i X_{it} + u_t \qquad (۶-۲)$$

با تخمین این مدل، R^2 نیز به‌دست می‌آید. از طرف دیگر، به‌عنوان مثال برای X_i داریم:

$$X_{it} = \gamma_1 + \gamma_2 X_{2t} + \gamma_3 X_{3t} + \cdots + \gamma_K X_K + \varepsilon_t = \gamma_1 + \sum_{\substack{j=1 \\ j \neq i}}^{K} \gamma_j X_{jt} + \varepsilon_t \qquad (۶-۳)$$

ضریب تعیین این معادله را با R_i^2 نشان داده و با ضریب تعیین مدل اصلی(R^2) مقایسه می‌کنیم. اگر $R_i^2 > R^2$ باشد در این صورت همخطی شدید است. اما این روش به سختی می‌تواند در تشخیص همخطی کمک کند، زیرا تصور کنید که $R^2 = ۰/۹۴$ و $R_i^2 = ۰/۹۲$ باشد در این صورت رابطه خطی شدیدی بین X_i و سایر Xها وجود دارد، هر چند که R_i^2 اندکی از R^2 کوچکتر است. بنابراین روش مناسب‌تر آن است که معنادار بودن R_i^2 را آزمون کنیم. بدین صورت که اگر R_i^2 معنی‌دار باشد بیانگر وجود همخطی شدید بین X_i با سایر Xها است. برای آزمون معنی‌دار بودن R_i^2، از تابع F استفاده می‌کنیم:

$$F_i = \frac{ESS/(n-(K-1))}{RSS/((K-1)-1)} = \frac{n-K+1}{K-2}\frac{R_i^2}{1-R_i^2} \qquad (۶-۴)$$

درجه آزادی F_i به ترتیب $K-2$ و $n-K+1$ می‌باشد که n تعداد مشاهدات و K تعداد پارامترهای تخمینی در معادله رگرسیون X_i روی Xها می‌باشد. اگر مقدار F_i از F جدول بزرگتر باشد در این صورت R_i^2 به اندازه کافی بزرگ است و حاکی از همخطی شدید X_i با سایر Xها می‌باشد.

روش فوق هر چند مفید و ساده است اما ممکن است در شناسایی برخی از انواع همخطی نتواند کمک نماید. به‌عنوان مثال فرض کنید که ترکیب خطی از X_2 و X_3 دارای همخطی شدیدی با ترکیب خطی از سایر متغیرها باشد. در این صورت بایستی ترکیب خطی از X_2 و X_3 را روی سایر متغیرهای توضیحی برازش کنیم. در چنین شرایطی نمی‌توان از روش فوق استفاده نمود.

روش دیگر در شناسایی همخطی، بررسی دترمینان ماتریس $\mathbf{X'X}$ است. هرچه دترمینان $\mathbf{X'X}$ نزدیک به صفر باشد بیانگر هم‌خطی شدید بین X_i ها است، اما مشکل این روش این است که

دترمینان $\mathbf{X'X}$ چقدر باید به صفر نزدیک باشد تا بتوان گفت که بین X_i ها همخطی شدید وجود دارد.

۴-۲-۶ راه‌های کاهش همخطی

مشکل همخطی، مربوط به نمونه است. به‌عنوان مثال تصور کنید که مصرف را تابعی از درآمد و ثروت می‌دانیم. از قبل مشخص است که بین درآمد و ثروت یک رابطه مستقیم وجود دارد. یعنی افراد ثروتمندتر دارای درآمد بالاتر می‌باشند. اما نمی‌توان به این قضاوت اکتفا نمود، بلکه بایستی ببینیم که در نمونه یا بر اساس داده‌های جمع‌آوری شده آیا همخطی شدیدی بین درآمد و ثروت وجود دارد یا نه. لذا در بررسی همخطی، بایستی راجع به داده‌ها بحث نمود و نه در مورد مدل یا روش تخمین.

بر اساس بحث فوق یکی از راه‌های کاهش همخطی این است که حجم نمونه را افزایش دهیم. با افزایش حجم نمونه ممکن است همخطی کاهش یابد. به‌عنوان مثال در معادله‌ای که شامل دو متغیر X_2 و X_3 است، اگر با افزایش حجم نمونه ضریب همبستگی X_2 و X_3 تغییر نکند، در این صورت واریانس تخمین‌زننده‌ها کاهش می‌یابد. زیرا به‌عنوان مثال واریانس $\hat{\beta}_2$ برابر است با:

$$\text{var}(\hat{\beta}_2)=\frac{\sigma^2}{(1-r_{23}^2)\sum x_{i2}^2} \qquad (۶-۵)$$

با افزایش حجم نمونه، $\sum x_{i2}^2$ افزایش می‌یابد و لذا واریانس $\hat{\beta}_2$ را کاهش می‌دهد. با کاهش واریانس $\hat{\beta}_2$، امکان برآورد دقیق‌تری از این پارامتر فراهم می‌شود.

روش دیگر برای کاهش همخطی، حذف متغیری است که موجب همخطی می‌شود. اما مشکل این روش آن است که مدلی را که از نظر تئوری درست است تبدیل به مدلی می‌کنیم که از تئوری فاصله می‌گیرد. در چنین شرایطی دچار مشکل «تورش تصریح» می‌شویم که موجب تورش ضرایب تخمینی می‌شود. (به فصل ۲ و ۳ مراجعه شود).

راه دیگر برای کاهش همخطی، استفاده از ترکیب داده‌های سری زمانی و مقطعی است. به‌عنوان مثال فرض کنید می‌خواهیم تابع تقاضای یک کالا را برآورد کنیم و متغیرهای توضیحی ما شامل قیمت و درآمد می‌باشد. در اینجا می‌توان ضریب مربوط به متغیر درآمد را با استفاده از داده‌های مقطعی (گروه‌های درآمدی) برآورد نمود.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (۶-۶)$$

Y_t مصرف کالای مورد نظر، X_2 و X_3 به ترتیب درآمد و قیمت می‌باشند. در این حالت ابتدا معادله $Y_t = \alpha_1 + \alpha_2 X_{2t}$ را با داده‌های مقطعی برآورد کرده و سپس α_2 را به‌عنوان تخمین β_2 در نظر می‌گیریم. توجه شود که در اینجا فرض بر این است که $\hat{\alpha}_2 = \hat{\beta}_2$ خواهد بود. سپس بر اساس $\hat{\beta}_2$، معادله رگرسیون را به‌صورت زیر می‌نویسیم:

$$Y_t = \beta_1 + \hat{\beta}_2 X_{2t} + \beta_3 X_{3t} + u_t$$
$$Y_t - \hat{\beta}_2 X_{2t} = \beta_1 + \beta_3 X_{3t} + u_t$$
$$Y_t^* = \beta_1 + \beta_3 X_{3t} + u_t \quad , \quad Y_t^* = Y_t - \hat{\beta}_2 X_{2t} \qquad (۶-۷)$$

حال می‌توان معادله آخر را برآورد نمود و ضرایب $\hat{\beta}_1$ و $\hat{\beta}_3$ را نیز به‌دست آورد. یکی دیگر از روش‌هایی که می‌تواند به کاهش همخطی کمک نماید استفاده از تئوری است. به‌عنوان مثال تابع تولید را به‌صورت زیر در نظر بگیرید:

$$LnY_t = \alpha + \beta_1 LnL_t + \beta_2 LnK_t + u_t \qquad (۶-۸)$$

فرض کنید از قبل می‌دانیم که $\hat{Y}_t^3$ است. در این صورت تابع تولید را با توجه به $\beta_1 = 1 - \beta_2$ صورت زیر می‌نویسیم:

$$LnY_t = \alpha + (1-\beta_2) LnL_t + \beta_2 LnK_t + u_t$$
$$LnY_t - LnL_t = \alpha + \beta_2 (LnK_t - LnL_t) + u_t$$
$$Ln\left(\frac{Y_t}{L_t}\right) = \alpha + \beta_2 Ln\left(\frac{K_t}{L_t}\right) + u_t \qquad (۶-۹)$$

در چنین شرایطی اگر L و K همخطی شدید داشته باشند، با اعمال قید $\beta_1 + \beta_2 = 1$ تعداد متغیرهای توضیحی را تقلیل می‌دهیم. در واقع این راهی برای اعمال و آزمون قید $\beta_1 + \beta_2 = 1$ است، هر چند که به برطرف کردن همخطی نیز کمک می‌کند.

یکی دیگر از روش‌های کاهش همخطی، تبدیل متغیرها است. به‌عنوان مثال در برآورد تابع تولید، اگر به‌جای مقدار متغیرها از تفاضل آنها یا نرخ رشد آنها استفاده کنیم، معمولاً همخطی کاهش می‌یابد. تابع تولید کاب- داگلاس را درنظر بگیرید که لگاریتم آن به‌صورت زیر است:

$$LnY_t = \alpha + \beta_1 LnL_t + \beta_2 LnK_t + u_t$$

با دیفرانسیل گیری از تابع تولید، مدل زیر به‌دست می‌آید:

$$\dot{Y}_t = \beta_1 \dot{L}_t + \beta_2 \dot{K}_t + \varepsilon_t \quad (6\text{-}10)$$

$\dot{X}_t$ بیانگر نرخ رشد متغیر X_t می‌باشد. در اینجا همخطی بین $\dot{L}$ و $\dot{K}$ عملاً ناچیز است. اما مشکل این است که ε_t بیانگر تغییرات u_t است که به‌صورت $\varepsilon_t = \frac{du_t}{u_t} \cong \frac{u_t - u_{t-1}}{u_{t-1}}$ می‌باشد. در چنین شرایطی معمولاً ε_t دچار خودهمبستگی می‌شود که باید برای آن نیز چاره‌ای اندیشید.

۳-۶ فرم تابعی نادرست

یکی از فروض معادلات رگرسیون کلاسیک این است که فرم تابعی به‌درستی انتخاب شده است. اما ممکن است این فرض درست نباشد. برای آزمون خطی بودن مدل، رمزی (۱۹۶۹) آزمونی را ارائه نموده است که براساس آن می‌توان نادرست بودن شکل تابع را تشخیص داد. این آزمون معروف به آزمون RESETرمزی است. این آزمون از یک رگرسیون کمکی استفاده می‌کند. در این رگرسیون مقادیر باقیمانده ($\hat{u}_t$) بر روی توان‌های مختلف $\hat{Y}_t$ برازش می‌شود.

$$\hat{u}_t = \alpha_1 + \alpha_2 \hat{Y}_t^2 + \alpha_3 \hat{Y}_t^3 + \cdots + \alpha_p \hat{Y}_t^p + v_t \quad (6\text{-}11)$$

وقتی که $\hat{Y}_t$ با توان‌های مختلف در نظر گرفته می‌شود بدان معنا است که انواع روابط غیرخطی را در نظر گرفته‌ایم. به‌عنوان مثال $\hat{Y}_t^2$ عبارت است از:

$$\hat{Y}_t^2 = (\hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \cdots + \hat{\beta}_K X_{Kt})^2 \quad (6\text{-}12)$$

در اینجا وقتی پرانتز را به توان برسانیم، هم توان‌های دوم متغیرها و هم حاصل‌ضرب‌های آنها در نظر گرفته شده است.

برای آزمون درستی شکل معادله، از R^2 استفاده می‌کنیم. بدین منظور مقدار R^2 را برای مدل (۱۱-۶) به‌دست می‌آوریم. همان‌طور که قبلاً اشاره شد، nR^2 (برای معادله ۱۱-۶) دارای توزیع مجانبی χ^2 با درجه آزادی $p-1$ می‌باشد (p تعداد پارامترهای تخمینی است)[1]. اگر مقدار nR^2 از

---

[1] در فصل هفتم ثابت می‌شود که nR^2 برابر با آماره ضریب لاگرانژ (LM) است.

مقدار $\chi^۲$ جدول ($\chi^۲_{/.۰۵,p-۱}$) بزرگتر باشد، فرضیه صفر مبنی بر اینکه «فرم تابعی درست است» رد می‌شود.

آزمون RESETرمزی را با کمک تابع F نیز می‌توان انجام داد. بدین منظور فرض کنید که مدل مورد نظر به‌صورت (۱۳-۶) باشد:

$$Y_t = \beta_۱ + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + \cdots + \beta_K X_{Kt} + u_t \qquad (۶-۱۳)$$

با تخمین این مدل، مقدار $R^۲$ را محاسبه کرده و آن را با $R^۲_{old}$ نشان می‌دهیم. حال معادله (۱۳-۶) را در حالی که $\hat{Y}_t^۲$، $\hat{Y}_t^۳$ و ... را به آن اضافه شده است، برآورد می‌کنیم:

$$Y_t = \beta_۱ + \beta_۲ x_{۲t} + \cdots + \beta_K X_{Kt} + \gamma_۱ \hat{Y}_t^۲ + \gamma_۲ \hat{Y}_t^۳ + \cdots + \gamma_p \hat{Y}_t^{p+۱} + v_t \qquad (۶-۱۴)$$

با برآورد این مدل، مقدار $R^۲$ نیز به‌دست می‌آید که آن را $R^۲_{new}$ می‌نامیم. برای آزمون معنی‌دار بودن متغیرهای $\hat{Y}_t^۲$، $\hat{Y}_t^۳$ و ...، باید دید که آیا وارد نمودن آنها تفاوت محسوسی در قدرت توضیحی مدل ایجاد کرده است یا نه. بدیهی است که اگر $R^۲_{new}$ به اندازه کافی از $R^۲_{old}$ بزرگتر باشد می‌توان گفت که فرم تابعی (۱۳-۶) نادرست است و باید فرم دیگری را در نظر گرفت.

بدین منظور از تابع F استفاده می‌کنیم:

$$F_{p,n-K-p} = \frac{(R^۲_{new} - R^۲_{old})/p}{(۱-R^۲_{new})/(n-K-p)} = \frac{n-K-p}{p} \frac{R^۲_{new} - R^۲_{old}}{۱-R^۲_{new}} \qquad (۶-۱۵)$$

اگر مقدار F از F جدول ($F_{/.۰۵,p,n-K-p}$) بیشتر باشد، نشان می‌دهد که فرم تابعی (۱۳-۶) نادرست است و باید فرم تابعی دیگری را در نظر بگیریم. این فرم جدید ممکن است همراه با توانهای بالاتر برای Xها یا به‌صورت حاصل‌ضرب Xها باشد. همچنین ممکن است به‌صورت لگاریتمی و یا هر فرم غیرخطی دیگری باشد.

آزمون RESET رمزی در Eviews

فایل data2

با برآورد رابطه بین نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) و نرخ رشد نیروی کار (GL) و نرخ رشد موجـود سرمایه (GK)، نتایج زیر به‌دست آمده است:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 01:10
Sample (adjusted): 1346 1380
Included observations: 35 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.107676	0.174091	6.362644	0.0000
GK	0.418728	0.072549	5.771668	0.0000

R-squared	0.691456	Mean dependent var	5.677239
Adjusted R-squared	0.682106	S.D. dependent var	7.492675
S.E. of regression	4.224528	Akaike info criterion	5.775138
Sum squared resid	588.9391	Schwarz criterion	5.864015
Log likelihood	-99.06491	Hannan-Quinn criter.	5.805818
Durbin-Watson stat	1.578005		

در پنجره فوق، برای انجام آزمون RESET رمزی به‌صورت زیر عمل می‌کنیم:

View → Stability Tests → Ramsey RESET Test

در اینجا پنجره‌ای باز می‌شود که در مقابل گزینه Number of Fitted Terms عددی را باید وارد کنیم که بیانگر مقدار p یعنی توان چند جمله‌ای در معادله (١١-٦) می‌باشد. اگر عدد ١ را وارد کنیم بدان معنا است که فقط $\hat{Y}_t^2$ را در نظر گرفته‌ایم.

RESET Specification

Number of fitted terms:

OK | Cancel

با انتخاب OK، نتایج به‌صورت زیر نشان داده می‌شود:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Ramsey RESET Test
Equation: UNTITLED
Specification: GYNO GL GK
Omitted Variables: Squares of fitted values

	Value	df	Probability
t-statistic	1.902919	32	0.0661
F-statistic	3.621101	(1, 32)	0.0661
Likelihood ratio	3.752080	1	0.0527

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	59.86923	1	59.86923
Restricted SSR	588.9391	33	17.84664
Unrestricted SSR	529.0699	32	16.53343
Unrestricted SSR	529.0699	32	16.53343

LR test summary:

	Value	df
Restricted LogL	-99.06491	33
Unrestricted LogL	-97.18887	32

Unrestricted Test Equation:
Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 01:12
Sample: 1346 1380
Included observations: 35

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	0.871067	0.208657	4.174635	0.0002
GK	0.379099	0.072868	5.202536	0.0000
FITTED^2	0.016770	0.008813	1.902919	0.0661

R-squared	0.722821	Mean dependent var	5.677239
Adjusted R-squared	0.705497	S.D. dependent var	7.492675
S.E. of regression	4.066133	Akaike info criterion	5.725078
Sum squared resid	529.0699	Schwarz criterion	5.858394
Log likelihood	-97.18887	Hannan-Quinn criter.	5.771099
Durbin-Watson stat	1.497463		

در پنجره فوق، مقدار t مربوط به ضریب $\hat{Y}_t^{۲}$ (که در اینجا با FITTED^2 نشان داده می‌شود) است که در قسمت پایین نیز ارائه شده است. اما مقدار F و احتمال مربوطه نشان می‌دهد که فرم تابعی درست نمی‌باشد و نیاز به در نظر گرفتن فرم دیگری داریم. لازم به ذکر است که مقدار F از RSSهای مقید و غیر مقید به‌دست آمده است که در قسمت F-Test Summary ارائه شده‌اند. همچنین، بعد از F مقدار نسبت درستنمایی نیز محاسبه شده که برابر با ۳/۷۵ می‌باشد و با توجه به احتمال مربوطه، نشان می‌دهد که فرم تابعی، غلط می‌باشد. نسبت درستنمایی (LR) به‌صورت زیر به‌دست می‌آید که اجزای آن در قسمت LR test summary ارائه شده است:

$$LR=-2[\log(L_R)-\log(L_{UR})]= -۲[(-۹۹/۰٦٤۹۱)-(-۹۷/۱۸۸۸۷)] = ۳/۷۵۲۰۸۰$$

که L_R تابع درستنمایی مقید و L_{UR} تابع درستنمایی غیرمقید می‌باشد.

از طرف دیگر اگر FITTED^2 را که برابر با $(\hat{\beta}_۱ GL+\hat{\beta}_۲ GK)^۲$ است به توان ۲ برسانیم، متغیرهای توضیحی هم با توان ۲ و هم به‌صورت حاصل‌ضرب وارد معادله اصلی می‌شوند. با وارد نمودن این متغیرها، تخمین مدل عبارت است از:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 02:07
Sample (adjusted): 1346 1380
Included observations: 35 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.104422	0.396952	2.782256	0.0092
GK	0.282346	0.131255	2.151117	0.0396
GL^2	0.000903	0.040694	0.022197	0.9824
GK^2	0.011605	0.008633	1.344245	0.1889
GL*GK	0.002031	0.018039	0.112590	0.9111

R-squared	0.732655	Mean dependent var	5.677239
Adjusted R-squared	0.697009	S.D. dependent var	7.492675
S.E. of regression	4.124314	Akaike info criterion	5.803240
Sum squared resid	510.2990	Schwarz criterion	6.025433
Log likelihood	-96.55670	Hannan-Quinn criter.	5.879941
Durbin-Watson stat	1.668576		

جدول فوق نشان می‌دهد که ضرایب GL^2 و GK*GL کاملاً بی‌معنی هستند، لذا این دو را از معادله حذف کرده و فقط GK^2 را نگه می‌داریم. در این حالت، نتایج زیر به‌دست می‌آید:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 02:09
Sample (adjusted): 1346 1380
Included observations: 35 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.121350	0.164715	6.807802	0.0000
GK	0.277610	0.093577	2.966645	0.0057
GK^2	0.012176	0.005492	2.217030	0.0338

R-squared	0.732538	Mean dependent var	5.677239
Adjusted R-squared	0.715822	S.D. dependent var	7.492675
S.E. of regression	3.994224	Akaike info criterion	5.689392
Sum squared resid	510.5225	Schwarz criterion	5.822708
Log likelihood	-96.56437	Hannan-Quinn criter.	5.735413
Durbin-Watson stat	1.684315		

ملاحظه می‌شود که علاوه بر ضرایب GL و GK، ضریب GK^2 نیز معنی‌دار است. همچنین نسبت به مدل اصلی که فقط شامل متغیرهای توضیحی GL و GK است، دارای $\overline{R}^2$ بالاتری است.

۴-۶ حذف متغیرهای مهم

فرض کنید مدل صحیح به‌صورت زیر باشد:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (۶-۱۶)$$

حال تصور کنید که متغیر مهمی مانند X_{3t} را حذف کرده و مدل زیر را برآورد نماییم:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + u_t \qquad (۶-۱۷)$$

پیامد حذف متغیر مهم X_{3t} آن است که اولاً β_3 را صفر فرض کرده‌ایم و ثانیاً در برآورد ضریب X_{2t} دچار تورش می‌شویم، زیرا $E(\hat{\alpha}_2) \neq \beta_2$ است. تخمین ضریب X_{2t} از معادله (۱۷-۶) عبارت است از:

$$\hat{\alpha}_2 = \frac{\sum x_{2t} y_t}{\sum x_{2t}^2}$$

از معادله (۱۶-۶) به‌جای Y_t قرار داده و $\hat{\alpha}_2$ را به صورت زیر می‌نویسیم:

$$\hat{\alpha}_2 = \frac{\sum x_{2t} y_t}{\sum x_{2t}^2} = \frac{\sum x_{2t} Y_t}{\sum x_{2t}^2} = \frac{\sum x_{2t} (\beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t)}{\sum x_{2t}^2}$$

$$= \beta_2 + \beta_3 b_{32} + \frac{\sum x_{2t} u_t}{\sum x_{2t}^2}$$

زیرا $\sum x_{2t} = 0$ و $\sum x_{2t} X_{2t} = \sum x_{2t}^2$ است. همچنین b_{32} از رگرسیون $X_{3t} = b_1 + b_{32} X_{2t} + \varepsilon_t$ به‌دست می‌آید که برابر است با:

$$b_{32} = \frac{\sum x_{2t} x_{3t}}{\sum x_{2t}^2}$$

امید ریاضی $\hat{\alpha}_2$ برابر است با:

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 b_{32} + E\left(\frac{\sum x_{2t} u_t}{\sum x_{2t}^2}\right)$$

چون امید ریاضی u_t برابر با صفر است لذا خواهیم داشت:

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 b_{32} \qquad (۶-۱۸)$$

بنابراین مقدار تورش $\hat{\alpha}_2$ بستگی به β_3 و b_{32} دارد. اگر β_3 برابر با صفر باشد مقدار تورش نیز صفر خواهد شد. اما چنین نخواهد بود زیرا X_{3t} یک متغیر مهم است و لذا ضریب آن از نظر آماری معنادار است. حالت دیگری که تورش $\hat{\alpha}_2$ را صفر می‌کند، این است که b_{32} صفر باشد. اگر b_{32} صفر باشد بدین معنا است که متغیرهای توضیحی X_{2t} و X_{3t} هیچ رابطه‌ای ندارند و لذا برآورد ضریب X_{2t} متأثر از X_{3t} نخواهد بود.

در حالت کلی فرض کنید که شکل ماتریسی مدل صحیح به‌صورت زیر باشد:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad (۶-۱۹)$$

اما زیر برآورد می‌شود که $K-p$ متغیر از آن حذف شده است:

$$\mathbf{y} = \mathbf{X}_p \boldsymbol{\beta}_p + \mathbf{v} \qquad (۶\text{-}۲۰)$$

مـاتریس $\mathbf{X}_p$ شـامل متغیرهـای X_{1t} تـا X_{pt} مـی‌باشـد، در حـالی کـه $\mathbf{X}$ شـامل متغیرهـای X_{1t} تا X_{Kt} است که $K>p$ می‌باشد. بدین ترتیب متغیرهای X_{p+1} تا X_K از مدل حذف شده‌اند. با برآورد مدل (۶-۲۰)، تخمین β_p عبارت است از:

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}'_p \mathbf{X}_p)^{-1}(\mathbf{X}'_p \mathbf{y})$$

حال به‌جای $\mathbf{y}$ از (۶-۱۹) د (۶-۲۱) قرار می‌دهیم:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_p &= (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p (\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{u} \qquad (۶\text{-}۲۲)\end{aligned}$$

اما می‌توان ماتریس X را به‌صورت زیر نوشت:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} & X_{p+1,1} & X_{p+2,1} & \cdots & X_{K1} \\ X_{11} & X_{11} & \cdots & X_{p1} & X_{p+1,1} & X_{p+2,1} & \cdots & X_{K1} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ X_{11} & X_{11} & \cdots & X_{p1} & X_{p+2,1} & X_{p+2,1} & \cdots & X_{K1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_p & \mathbf{X}_E \end{bmatrix}$$

$\mathbf{X}_E$ ماتریسی است که شامل داده‌های مربوط به $\mathbf{X}_{p+1}$ تا $\mathbf{X}_K$ می‌باشد.

همچنین می‌تـوان $\boldsymbol{\beta}$ را بـه‌صـورت $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_p \\ \boldsymbol{\beta}_E \end{bmatrix}$ نوشـت کـه $\boldsymbol{\beta}_p$ بـرداری بـا ابعـاد $p \times 1$ و $\boldsymbol{\beta}_E$ برداری بـا ابعـاد $(K-p) \times 1$ مـی‌باشـد. بـا جایگـذاری بـه‌جـای $\mathbf{X}$ و $\boldsymbol{\beta}$ در (۶-۲۲) مـی‌تـوان $\hat{\boldsymbol{\beta}}_p$ را به‌صورت زیر نوشت:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_p &= (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \begin{bmatrix} \mathbf{X}_p & \mathbf{X}_E \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_p \\ \boldsymbol{\beta}_E \end{bmatrix} + (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{u} \\ &= (\mathbf{X}'_p \mathbf{X}_p)^{-1} \ (\mathbf{X}'_p \mathbf{X}_p \beta_p \mathbf{X}'_p \mathbf{X}_E \boldsymbol{\beta}_E) + (\mathbf{X}'_p \mathbf{X}_p)^{-1} (\mathbf{X}'_p \mathbf{u}) \\ &= \boldsymbol{\beta}_p + (\mathbf{X}'_p \mathbf{X}_p)^{-1} (\mathbf{X}'_p \mathbf{X}_E) \boldsymbol{\beta}_E + (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{u}\end{aligned}$$

امید ریاضی $\hat{\boldsymbol{\beta}}_p$ عبارت است از:

$$E(\hat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + (\mathbf{X}'_p \mathbf{X}_p)^{-1} (\mathbf{X}'_p \mathbf{X}_E) \boldsymbol{\beta}_E \qquad (۶\text{-}۲۳)$$

توجه شود که $(\mathbf{X}'_p \mathbf{X}_p)^{-1}(\mathbf{X}'_p \mathbf{X}_E)$ بیانگر ضرایب رگرسیون K-p متغیر حذف‌شـده ازمـدل (۶-۱۶) روی p متغیر X موجود در مدل (۶-۲۰) می‌باشد. بدین منظور، ضرایب آنها را با مـاتریس $\mathbf{b}$ نشان می‌دهیم $(\mathbf{X}_E = \mathbf{X}_p \mathbf{b} + \mathbf{v})$ که $\hat{\mathbf{b}} = (\mathbf{X}'_p \mathbf{X}_p)^{-1} (\mathbf{X}'_p \mathbf{X}_E)$ می‌باشد. بنابراین $\mathbf{b}$ ماتریسی بـا ابعاد $p \times (K-p)$ می‌باشد که عبارت است از:

$$\mathbf{b}=\begin{bmatrix} b_{1,p+1} & b_{1,p+2} & \cdots & b_{2,K} \\ b_{2,p+1} & b_{2,p+2} & \cdots & b_{p,K} \\ \vdots & \vdots & \cdots & \vdots \\ b_{p,p+1} & b_{p,p+2} & \cdots & b_{p,K} \end{bmatrix}_{p\times(K-p)}$$

بنابراین می‌توان (۲۳-۶) را به‌صورت زیر نوشت:

$$E(\hat{\boldsymbol{\beta}}_p)=\boldsymbol{\beta}_p+\mathbf{b}\boldsymbol{\beta}_E=\begin{bmatrix}\mathbf{I}_2 & \mathbf{b}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_p \\ \boldsymbol{\beta}_E\end{bmatrix}=\begin{bmatrix}\mathbf{I}_2 & \mathbf{b}\end{bmatrix}\boldsymbol{\beta}$$

که $\mathbf{I}_2=\begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}$ می‌باشد. سطر iام رابطه فوق عبارت است از:

$$E(\hat{\beta}_{pi})=\beta_i+b_{i,p+1}\beta_{p+1}+b_{i,p+1}\beta_{p+2}+\cdots+b_{i,K}\beta_K \quad , \quad i=1,2,\ldots,p \qquad (۶-۲۴)$$

بدین ترتیب مشاهده می‌شود که تخمین پارامترهایی که از معادله (۲۰-۶) به‌دست می‌آید، بدون تورش نمی‌باشند، زیرا $E(\hat{\beta}_{pi})\neq\beta_i$ می‌باشد. فقط در صورتی این تخمین‌ها بدون تورش هستند که عناصر ماتریس b برابر با صفر باشند، یعنی متغیرهای حذف شده با متغیرهایی که در مدل مانده‌اند هیچ‌گونه رابطه خطی نداشته باشند. علاوه بر این، حذف متغیرهای مهم باعث می‌شود که برآورد واریانس رگرسیون (σ^2) نیز دارای تورش گردد. بدین ترتیب، تخمین انحراف معیار ضرایب نیز تورش خواهند داشت که موجب می‌شود تا نتیجه‌گیری‌هایی که در مورد پارامترها صورت می‌گیرد، دچار خدشه شوند.

اما می‌توان تأثیر «متغیر حذف‌شده» را آزمون نمود. البته این آزمون تا حدود زیادی مشابه با آزمون محدودیت‌ها است که در فصل سوم بحث شد. به‌عنوان مثال فرض کنید که Y تابعی از دو متغیر X_2 و X_3 است، ولی X_3 از مدل حذف شده است. در این حالت ضریب X_3 را برابر با صفر قرار داده‌ایم ($\beta_3=0$).

$$Y_t=\beta_1+\beta_2X_{2t}+\beta_3X_{3t}+u_t \qquad (۶-۲۵)$$

$$Y_t=\alpha_1+\alpha_2X_{2t}+v_t \qquad (۶-۲۶)$$

معادله (۲۵-۶) و (۲۶-۶) به‌ترتیب رگرسیون‌های غیرمقید و مقید را نشان می‌دهد که ضریب تعیین آنها به‌ترتیب R^2_{UR} و R^2_R می‌باشد. برای آزمون این فرضیه، از F استفاده می‌کنیم که عبارت است از:

$$F_{1,n-K}=(n-K)\frac{R^2_{UR}-R^2_R}{1-R^2_{UR}} \qquad (۶-۲۷)$$

توجه داریم که اگر در توزیع F، درجه‌آزادی صورت برابر ۱ باشد، آنگاه F برابر با t^2 است.

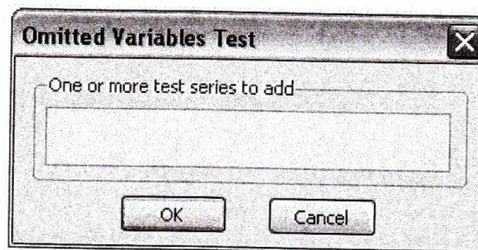
آزمون متغیرهای حذف شده با استفاده از Eviews

فایل data2

ابتدا مدل مورد نظر را که به‌عنوان مثال معادله (۲۶-۶) می‌باشد، برآورد می‌کنیم که نتایج آن در پنجره Equation نشان داده می‌شود. برای بررسی اثر متغیر حذف شده، به‌صورت زیر عمل می‌کنیم:

View → Coefficient Diagnostics → Omitted Variable Test

در این حالت پنجره‌ای به نام Omitted–Redundant Variable Test باز می‌شود، که در آن می‌توان نام متغیر یا متغیرهای حذف شده را وارد نمود.



بعد از وارد نمودن نام متغیرهای حذف شده (مثلاً X_3) و انتخاب OK، نتایج حاصل از تخمین مدل جدید (یعنی مدلی که به آن متغیرهای حذف شده، اضافه شده است) ارائه می‌شود. علاوه بر تخمین مدل جدید، برای آزمون فرضیهٔ متغیرهای حذف شده، مقادیر F و نسبت درست‌نمایی نیز داده می‌شود. اگر مقدار F بزرگتر از F جدول باشد یا مقادیر احتمال که در مقابل F نوشته می‌شود کمتر از ۰/۰۵ باشد، در این صورت متغیرهای حذف شده اثر معنی‌داری داشته‌اند.

به‌عنوان مثال در آزمون Resetرمزی دیدیم که مدل صحیح برای تبیین نرخ رشد تولید ناخالص داخلی به‌گونه‌ای است که بایستی توان دوم نرخ رشد سرمایه را نیز وارد کنیم. این موضوع را می‌توان از طریق آزمون متغیرهای حذف شده نیز بررسی نمود. بدین منظور ابتدا نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) را روی نرخ رشد نیروی کار (GL) و نرخ رشد سرمایه (GK) برازش می‌کنیم:

Equation: UNTITLED Workfile: YNO::Untitled\

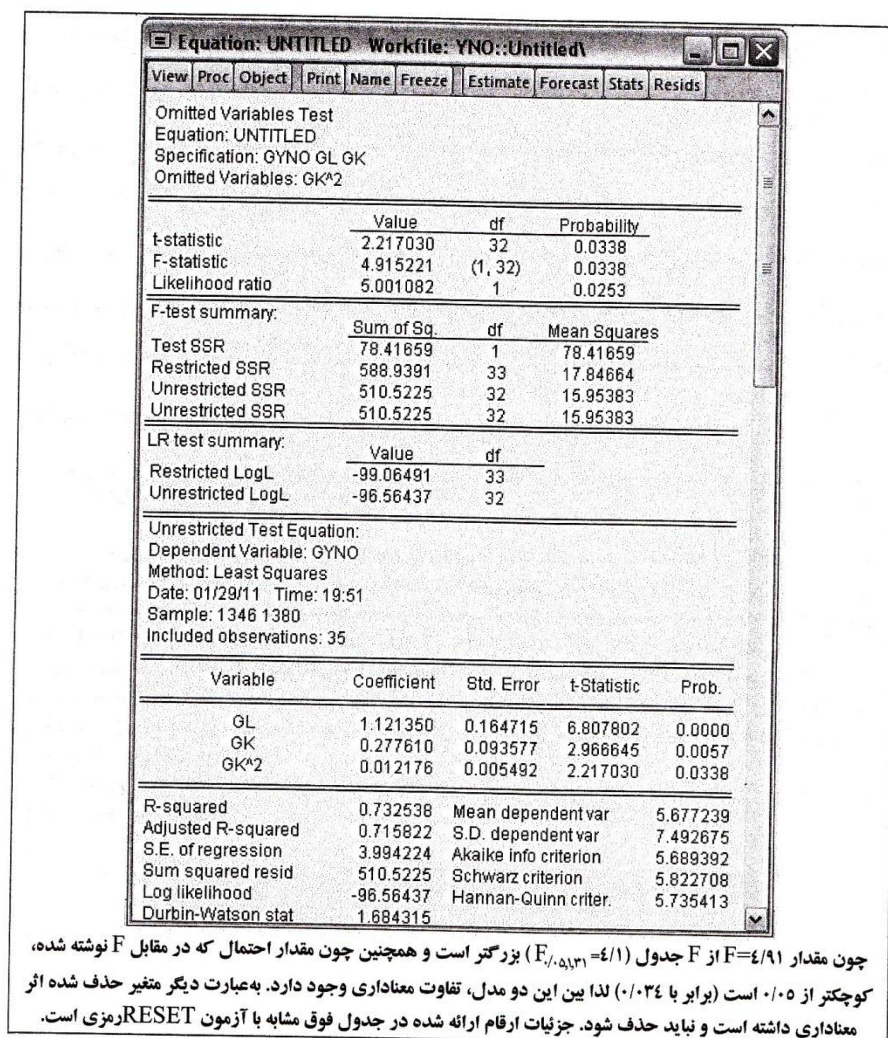
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 19:47
Sample (adjusted): 1346 1380
Included observations: 35 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.107676	0.174091	6.362644	0.0000
GK	0.418728	0.072549	5.771668	0.0000

R-squared	0.691456	Mean dependent var	5.677239
Adjusted R-squared	0.682106	S.D. dependent var	7.492675
S.E. of regression	4.224528	Akaike info criterion	5.775138
Sum squared resid	588.9391	Schwarz criterion	5.864015
Log likelihood	-99.06491	Hannan-Quinn criter.	5.805818
Durbin-Watson stat	1.578005		

در پنجرهٔ فوق، به طریقی که گفته شد اثر متغیر حذف شده (یعنی توان دوم نرخ رشد سرمایه) را بررسی می‌کنیم. نتایج حاصله عبارت است از:



Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Omitted Variables Test
Equation: UNTITLED
Specification: GYNO GL GK
Omitted Variables: GK^2

	Value	df	Probability
t-statistic	2.217030	32	0.0338
F-statistic	4.915221	(1, 32)	0.0338
Likelihood ratio	5.001082	1	0.0253

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	78.41659	1	78.41659
Restricted SSR	588.9391	33	17.84664
Unrestricted SSR	510.5225	32	15.95383
Unrestricted SSR	510.5225	32	15.95383

LR test summary:

	Value	df
Restricted LogL	-99.06491	33
Unrestricted LogL	-96.56437	32

Unrestricted Test Equation:
Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 19:51
Sample: 1346 1380
Included observations: 35

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.121350	0.164715	6.807802	0.0000
GK	0.277610	0.093577	2.966645	0.0057
GK^2	0.012176	0.005492	2.217030	0.0338

R-squared	0.732538	Mean dependent var	5.677239
Adjusted R-squared	0.715822	S.D. dependent var	7.492675
S.E. of regression	3.994224	Akaike info criterion	5.689392
Sum squared resid	510.5225	Schwarz criterion	5.822708
Log likelihood	-96.56437	Hannan-Quinn criter.	5.735413
Durbin-Watson stat	1.684315		

چون مقدار $F=٤/٩١$ از F جدول ($F_{/٠٥,١,٣١}=٤/١$) بزرگتر است و همچنین چون مقدار احتمال که در مقابل F نوشته شده، کوچکتر از ۰/۰۵ است (برابر با ۰/۰۳٤) لذا بین این دو مدل، تفاوت معناداری وجود دارد. به‌عبارت دیگر متغیر حذف شده اثر معناداری داشته است و نباید حذف شود. جزئیات ارقام ارائه شده در جدول فوق مشابه با آزمون RESETرمزی است.

٥-٦ ورود متغیرهای نامربوط

در اینجا بر خلاف حذف متغیرهای مهم، برخی متغیرهایی که ضرورتی ندارند وارد مدل شده‌اند. به‌عنوان مثال فرض کنید که مدل صحیح به‌صورت زیر باشد:

$$Y_t = \beta_۱ + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + \beta_۴ X_{۴t} + u_t \qquad (۶-۲۸)$$

اما به‌جای (۲۸-۶) مدل زیر را برآورد می‌کنیم که شامل متغیر نامربوط $X_۵$ می‌باشد:

$$Y_t = \beta_۱ + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + \beta_۴ X_{۴t} + \beta_۵ X_{۵t} + u_t \qquad (۶-۲۹)$$

از آنجا که X_5 یک متغیر نامربوط و غیرضروری است لذا انتظار داریم که β_5 برابر با صفر باشد. پیامد ورود چنین متغیری این است که سایر ضرایب هنوز سازگار و بدون تورش هستند. اما تخمین‌زننده‌ها را ناکارآ می‌کند و این موجب می‌شود که انحراف معیار تخمین‌زننده‌ها بزرگ شود. اما به‌طورکلی میزان کاهش در کارایی تخمین‌زننده‌ها بستگی به قدرمطلق ضریب همبستگی بین متغیر نامربوط و سایر متغیرهای توضیحی دارد. زیرا واریانس تخمین‌زننده‌ها رابطه مستقیم با مجذور ضریب همبستگی بین متغیرهای توضیحی دارد. به‌طور کلی اگر متغیر نامربوطی را وارد مدل کنیم که در عین حال با بقیه متغیرهای موجود در مدل ایجاد همخطی نماید، باعث کاهش کارایی تخمین‌زننده‌ها می‌گردد.

آزمون متغیرهای نامربوط با استفاده از Eviews فایل data2

مراحل انجام این آزمون دقیقاً مشابه با آزمون متغیرهای حذف شده است. ولی توجه داشته باشیم که در اینجا نام متغیرهایی را وارد می‌کنیم که بایستی از مدل حذف شوند. به‌عنوان مثال ابتدا مدل (۲۹-٦) را تخمین می‌زنیم و سپس در پنجره‌ای که نتایج تخمین نشان داده شده است از طریق View → Coefficient Diagnostics → Redundant Variable Test پنجره‌ای باز می‌شود که در آن، نام متغیرهای نامربوط را وارد می‌کنیم و با انتخاب OK نتایج مورد نظر به‌دست می‌آید. نتایج حاصل از برآورد مدل جدید به همراه مقدار F و احتمال آن نیز داده می‌شود. در اینجا نیز اگر مقدار F بزرگتر از F جدول و یا احتمال‌های آن کمتر از ۰/۰۵ باشد در این صورت متغیری را که از مدل اولیه کنار می‌گذاریم یک متغیر نامربوط و زائد نیست و بایستی در مدل حفظ شود.

به‌عنوان مثال معادله سرمایه‌گذاری خصوصی (IP) روی نرخ بهره حقیقی (RR) و درآمد ملی (Y) برازش شده است. نتیجه عبارت است از:

Equation: UNTITLED Workfile: S3::Untitled\
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

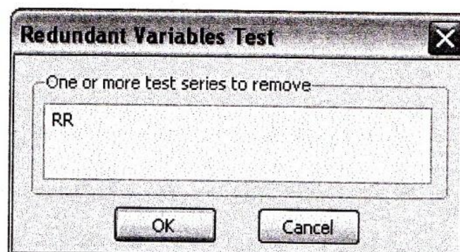
Dependent Variable: IP
Method: Least Squares
Date: 10/13/11 Time: 15:59
Sample: 1345 1380
Included observations: 36

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7386.982	5077.692	-1.454791	0.1552
RR	97.82949	165.7167	0.590342	0.5590
Y	0.159006	0.023502	6.765674	0.0000

R-squared	0.582922	Mean dependent var	25513.35
Adjusted R-squared	0.557645	S.D. dependent var	13257.30
S.E. of regression	8817.405	Akaike info criterion	21.08650
Sum squared resid	2.57E+09	Schwarz criterion	21.21846
Log likelihood	-376.5570	Hannan-Quinn criter.	21.13256
F-statistic	23.06096	Durbin-Watson stat	1.063540
Prob(F-statistic)	0.000001		

حال بررسی می‌کنیم که آیا نرخ بهره (RR) یک متغیر زائد است یا خیر. بدین منظور، در پنجره فوق، مسیر

View→Coefficient Diagnostics → Redundant Variable Test را انتخاب می‌کنیم که پنجره Redundant Variable Test باز می‌شود. در این پنجره نام متغیر مورد نظر (متغیر زائد) را وارد می‌کنیم.



با انتخاب OK نتایج به‌صورت زیر به‌دست می‌آید:

Equation: UNTITLED Workfile: S3::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Redundant Variables Test
Equation: UNTITLED
Specification: IP C RR Y
Redundant Variables: RR

	Value	df	Probability
t-statistic	0.590342	33	0.5590
F-statistic	0.348504	(1, 33)	0.5590
Likelihood ratio	0.378192	1	0.5386

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	27094977	1	27094977
Restricted SSR	2.59E+09	34	76256871
Unrestricted SSR	2.57E+09	33	77746626
Unrestricted SSR	2.57E+09	33	77746626

LR test summary:

	Value	df
Restricted LogL	-376.7461	34
Unrestricted LogL	-376.5570	33

Restricted Test Equation:
Dependent Variable: IP
Method: Least Squares
Date: 10/13/11 Time: 16:08
Sample: 1345 1380
Included observations: 36

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7369.472	5028.722	-1.465476	0.1520
Y	0.156605	0.022924	6.831371	0.0000

R-squared	0.578517	Mean dependent var	25513.35
Adjusted R-squared	0.566121	S.D. dependent var	13257.30
S.E. of regression	8732.518	Akaike info criterion	21.04145
Sum squared resid	2.59E+09	Schwarz criterion	21.12942
Log likelihood	-376.7461	Hannan-Quinn criter.	21.07215
F-statistic	46.66764	Durbin-Watson stat	1.073290
Prob(F-statistic)	0.000000		

در بالای جدول مقدار F داده شده است. چون مقدار F کوچک است (و یا احتمال مربوط به آن بزرگتر از ۰/۰۵ است) لذا زائد بودن RR رد نمی‌شود. در اینجا سایر مقادیری که داده شده است مشابه آزمون RESET رمزی است. در پایین این جدول، معادله‌ای برآورد شده است که بدون متغیر زائد می‌باشد.

۶-۶ آزمون‌های ثبات ضرایب

معادله رگرسیون زیر را در نظر بگیرید:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (۶-۳۰)$$

فرض ضمنی در این معادله این است که ضرایب β_1، β_2 و β_3 ثابت هستند. به‌عبارت دیگر این ضرایب چه در کل دوره مورد بررسی برآورد شوند و چه در بخشی از دوره مورد بررسی برآورد گردند، نتایج یکسانی به‌دست خواهد آمد. به‌عنوان مثال اگر معادله‌ای برای دوره ۸۰-۱۳۳۸ برآورد می‌کنیم ضرایب این معادله برای کل این دوره همان چیزی است که در دوره قبل و بعد از انقلاب به‌دست می‌آید.

اما این فرض را می‌توان در معرض آزمون قرار داد، زیرا ممکن است چنین فرضی برقرار نباشد. بدین منظور می‌توان کل دوره مورد مطالعه را به دو دوره و یا چند دوره تقسیم نمود و فرض ثبات ضرایب را مورد بررسی و آزمون قرار داد. یکی از آزمون‌های تغییر ضرایب، معروف به آزمون چاو است که در اینجا روش انجام این آزمون را بررسی می‌کنیم. روش دیگر، استفاده از متغیرهای مجازی است که در ادامه بررسی خواهد شد.

۱-۶-۶ آزمون نقطه شکست چاو[1]

مراحل انجام آزمون چاو عبارت است از:

۱- دوره مورد مطالعه را به دو دوره تقسیم می‌کنیم که دوره اول و دوم به‌ترتیب شامل n_1 و n_2 مشاهده می‌باشند ($n_1 + n_2 = n$).

۲- رگرسیون مورد نظر را برای دوره اول و دوره دوم برآورد می‌کنیم. مجموع مجذور خطا برای دوره اول را با RSS_1 و برای دوره دوم با RSS_2 نشان می‌دهیم و مجموع این‌دو را $RSS_{UR} = RSS_1 + RSS_2$ می‌نامیم که بیانگر رگرسیون غیرمقید است، زیرا ضرایب آزاد هستند تا در این دو دوره، متفاوت باشند. درجه آزادی RSS_1 و RSS_2 به‌ترتیب برابر با $n_1 - K$ و $n_2 - K$ می‌باشد و لذا درجه آزادی RSS_{UR} نیز برابر با $n-2K$ خواهد بود.

---

[1] Chow breakpoint test

۳- معادله مورد نظر را برای کل دوره برآورد کرده و مجموع مجذور باقیمانده‌های آن را با $RSS_R = RSS$ نشان می‌دهیم که بیانگر رگرسیون مقید است، زیرا ضرایب در هر دو دوره یکسان در نظر گرفته شده‌اند. درجه آزادی RSS_R برابر با $n-K$ است.

۴- برای بررسی تغییر ساختاری، بایستی RSS_{UR} و RSS_R را مقایسه کنیم. بدین منظور از نسبت F استفاده می‌کنیم:

$$F_{K,n-2K} = \frac{n-2K}{K}\frac{RSS_R - RSS_{UR}}{RSS_{UR}} = \frac{n-2K}{K}\frac{RSS-(RSS_1+RSS_2)}{RSS_1+RSS_2} \qquad (۶-۳۱)$$

تفاوت رگرسیون مقید و غیرمقید، فقط در ضرایب آنها است. اگر ضرایب باثبات باشند در این صورت تفاوت چندانی بین RSS و RSS_1+RSS_2 وجود ندارد و لذا مقدار F نیز کوچک می‌شود، در غیر این صورت مقدار F بزرگ شده و فرضیه ثبات ضرایب رد می‌شود.

آزمون نقطه شکست چاو در Eviews — فایل data2

تصور کنید که می‌خواهیم نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) را روی نرخ رشد کار (GL) و نرخ رشد سرمایه (GK) برازش کنیم. بدین منظور ابتدا معادله مورد نظر را برای دوره ۸۰-۱۳٤٦ برآورد می‌کنیم که حجم نمونه n=۳۵ می‌باشد. نتایج تخمین عبارت است از:

$$GYNO = 1.107GL + 0.419GK \qquad R^2 = 0.691 \quad \bar{R}^2 = 0.682$$
$$(6.4) \qquad (5.8) \qquad RSS = 588.94$$

توجه شود که در پنجرهٔ نتایج، RSS با Sum squared resid نشان داده شده است. حال می‌خواهیم ثبات ضرایب نیروی کار و سرمایه طی دوره مورد نظر را بررسی کنیم. بدین منظور دوره قبل و بعد از انقلاب را در نظر می‌گیریم. برای انجام آزمون چاو، در پنجرهٔ نتایج تخمین، به‌صورت زیر عمل می‌کنیم:

View → Stability Diagnostics → Chow Breakpoint Test

در پنجرهٔ Chow Test سال شروع دوره دوم را وارد کنیم (در این مثال سال ۱۳۵۷ می‌باشد):

Chow Tests

Enter one or more breakpoint dates

1357

Regressors to vary across breakpoints

gl gk

OK   Cancel

در پنجره فوق، متغیرهای توضیحی و همچنین عرض از مبدأ را می‌توان وارد نمود. اگر فقط C را بنویسیم، آنگاه فقط تغییر در عرض از مبدأ را لحاظ می‌کند. اما اگر C و GL را بنویسیم، آنگاه تغییر در عرض از مبدأ و ضریب GL را

لحاظ می‌کند. همچنین اگر تمام متغیرها را بنویسیم، در این صورت تغییر تمام ضرایب را لحاظ کرده و آزمون می‌کند.
با انتخاب OK، نتایج آزمون چاو به‌صورت زیر نشان داده می‌شود:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Chow Forecast Test
Equation: UNTITLED
Specification: GYNO GL GK
Test predictions for observations from 1357 to 1380

	Value	df	Probability
F-statistic	0.420224	(24, 9)	0.9565
Likelihood ratio	26.30943	24	0.3377

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	311.2160	24	12.96733
Restricted SSR	588.9391	33	17.84664
Unrestricted SSR	277.7231	9	30.85813
Unrestricted SSR	277.7231	9	30.85813

LR test summary:

	Value	df
Restricted LogL	-99.06491	33
Unrestricted LogL	-85.91019	9

Unrestricted log likelihood adjusts test equation results to account for observations in forecast sample

Unrestricted Test Equation:
Dependent Variable: GYNO
Method: Least Squares
Date: 01/29/11 Time: 20:51
Sample: 1346 1356
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GL	1.415792	0.465405	3.042062	0.0140
GK	0.481937	0.127818	3.770482	0.0044

R-squared	0.458899	Mean dependent var	12.18188
Adjusted R-squared	0.398777	S.D. dependent var	7.164186
S.E. of regression	5.555009	Akaike info criterion	6.430243
Sum squared resid	277.7231	Schwarz criterion	6.502587
Log likelihood	-33.36634	Hannan-Quinn criter.	6.384640
Durbin-Watson stat	1.152511		

با توجه به مقدار F که از F جدول کوچکتر است و همچنین احتمال آن که از ۰/۰۵ بزرگتر است، فرضیه ثبات ضرایب رد نمی‌شود. لذا ضرایب این معادله، در دوره‌های قبل و بعد از انقلاب، متفاوت نیست.

توجه شود که آزمون چاو را می‌توان برای چند دوره نیز انجام داد. به‌عنوان مثال اگر بخواهیم، آزمون ثبات ضرایب را برای دوره‌های ۵۶-۱۳۴۶، ۷۰-۱۳۵۷ و ۸۰-۱۳۷۱ انجام دهیم، لازم است که در پنجره مورد نظر، به ترتیب سال ۱۳۵۷ (سال شروع دوره دوم) و سال ۱۳۷۱ (سال شروع دوره سوم) را وارد نمائیم.

۲-۶-۶ آزمون پیش‌بینی چاو[1]

آزمون پیش‌بینی چاو بر این اساس است که اگر ضرایب باثبات باشند، خطای پیش‌بینی کوچک خواهد بود. بدین منظور کل داده‌ها را به دو دوره تقسیم می‌کنیم: دوره اول با n_1 مشاهده و دوره دوم با n_2 مشاهده. دوره اول را برای تخمین و دوره دوم را برای پیش‌بینی در نظر

[1] Chow forecast test

می‌گیریم. اگر ضرایب با ثبات باشند، بایستی متوسط خطای پیش‌بینی در دوره دوم (σ_f^2) و متوسط خطای رگرسیون در دوره اول (σ^2)، تقریباً یکسان باشند. بنابراین، فرضیه ثبات ضرایب به‌صورت $H_0 : \sigma^2 = \sigma_f^2$ می‌باشد. در غیر این صورت، رگرسیون دچار شکستگی شده است و لذا رگرسیون بر اساس داده‌های گذشته، نمی‌تواند مقادیر آینده را به‌خوبی پیش‌بینی کند. بدین منظور، مراحل زیر را انجام می‌دهیم:

۱- معادله $Y_t = \alpha + \beta X_t + u_t$ را برای n_1 مشاهدهٔ اول (دوره اول) برآورد کرده و مجموع مجذور باقیمانده را با RSS_1 نشان می‌دهیم.

$$RSS_1 = \sum_{t=1}^{n_1}(Y_t - \hat{Y}_t)^2 \quad , \quad \hat{\sigma}^2 = \frac{RSS_1}{n_1 - K} \qquad (۶-۳۲)$$

۲- برای دوره دوم که از سال $n_1 + 1$ تا n می‌باشد، مقادیر Y_t را با استفاده از رگرسیون برآوردی در مرحلهٔ ۱، پیش‌بینی می‌کنیم:

$$Y_t^f = \hat{\alpha} + \hat{\beta} X_t \quad ; \quad t = n_1 + 1, \ldots, n \qquad (۶-۳۳)$$

۳- خطاهای پیش‌بینی را به‌صورت زیر برای دوره دوم حساب می‌کنیم:

$$\begin{aligned} e_t^f &= Y_t - Y_t^f \\ &= (\alpha + \beta X_t + u_t) - (\hat{\alpha} + \hat{\beta} X_t) \\ &= -(\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_t + u_t \quad ; \quad t = n_1 + 1, \ldots, n \end{aligned} \qquad (۶-۳۴)$$

امید ریاضی و واریانس خطای پیش‌بینی عبارت است از (به بخش ۱۴-۲ مراجعه شود):

$$\begin{aligned} E(e_t^f) &= 0 \\ \text{var}(e_t^f) &= \sigma^2 \left(1 + \frac{1}{n} + (X_t - \overline{X})^2 \Big/ \sum_{t=n_1+1}^{n} x_t^2 \right) \end{aligned} \qquad (۶-۳۵)$$

همچنین مجموع مجذور خطاهای پیش‌بینی و میانگین خطای پیش‌بینی عبارت است از:

$$RSS_f = \sum_{t=n_1+1}^{n}(Y_t - Y_t^f)^2 \quad , \quad \hat{\sigma}_f^2 = \frac{RSS_f}{n_2} \qquad (۶-۳۶)$$

توجه شود که در اینجا، درجه آزادی برابر با n_2 است، زیرا برای محاسبه RSS_f هیچ ضریبی را برآورد نکرده‌ایم.

۴- در صورتی که ضرایب ثابت باشند نباید تفاوت معناداری بین $\hat{\sigma}_f^2$ و $\hat{\sigma}^2$ باشد. برای آزمون فرضیه $H_0 : \sigma^2 = \sigma_f^2$، آماره F را تشکیل می‌دهیم:

$$F_{n_2,n_1-k}=\frac{\chi^2_{n_2}/n_2}{\chi^2_{n_1-K}/(n_1-K)}=\frac{\hat{\sigma}^2_f/\sigma^2_f}{\hat{\sigma}^2/\sigma_2} \qquad (6\text{-}37)$$

تحت فرضیه H_0، آماره F عبارت است از:

$$F_{n_2,n_1-K}=\left.\frac{\hat{\sigma}^2_f/\sigma^2_f}{\hat{\sigma}^2/\sigma_2}\right|_{H_0}=\frac{\hat{\sigma}^2_f}{\hat{\sigma}^2} \qquad (6\text{-}38)$$
$$=\frac{RSS_f/n_2}{RSS_1/n_1-K}=\frac{n_1-K}{n_2}\frac{RSS_f}{RSS_1}$$

اگر فرضیه H_0 درست باشد، مقدار F کوچک خواهد بود و در ناحیه بحرانی قرار نمی‌گیرد و لذا ضرایب باثبات خواهند بود.

آزمون پیش‌بینی چاو در Eviews فایل data2

ابتدا معادله را برای کل دوره برآورد کرده و سپس مشابه آزمون چاو عمل می‌کنیم، با این تفاوت که به جای گزینه Chow Breakpoint Test گزینه Chow Forecast Test را انتخاب می‌کنیم. نتایج در پنجره‌ای به‌صورت زیر نشان داده می‌شود.

Equation: UNTITLED Workfile: AAA3::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Chow Forecast Test
Equation: UNTITLED
Specification: GYNO GL GK
Test predictions for observations from 1357 to 1380

	Value	df	Probability
F-statistic	0.309543	(24, 9)	0.9896
Likelihood ratio	21.06390	24	0.6350

چون مقدار F کوچک است و در ناحیه بحرانی قرار ندارد، لذا فرضیه ثبات ضرایب رد نمی‌شود.

۳-۶-۶ آزمون‌های بازگشتی

آزمون‌های بازگشتی مبتنی بر برآورد ضرایب برای دوره‌های مختلف است. لذا این روش‌ها حالت عمومی‌تری از آزمون چاو هستند. اگر دوره‌ها را با s نشان دهیم، دوره اول شامل K مشاهدهٔ اول (K تعداد متغیرهای توضیحی است)، دوره دوم شامل $K+1$ مشاهدهٔ بعدی و دوره آخر شامل کل مشاهدات(n) است. بنابراین، برآورد ضرایب رگرسیون ساده، عبارت است از:

$$\hat{\alpha}_s = \overline{Y}_s - \hat{\beta}_s \overline{X}_s \ , \ \hat{\beta}_s = \frac{\sum_{t=1}^{s} x_t y_t}{\sum_{t=1}^{s} x_t^2} \quad ; \quad s = K, K+1, \ldots, n \tag{۶-۳۹}$$

$$\overline{Y}_s = \frac{\sum_{t=1}^{s} Y_t}{s} \quad , \quad \overline{X}_s = \frac{\sum_{t=1}^{s} X_t}{s}$$

در حالت کلی (رگرسیون Kمتغیره) می‌توان تخمین‌های بازگشتی را به‌صورت زیر نشان داد:

$$\boldsymbol{\beta}_s = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{y}_s \quad ; \quad s = K, K+1, \cdots, n \tag{۶-۴۰}$$

از طرف دیگر به‌ازای هر یک از برآوردهای بازگشتی، RSS_s را محاسبه می‌کنیم:

$$RSS_s = \sum_{t=1}^{s} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{s} (Y_t - \hat{\alpha}_s - \hat{\beta}_s X_t)^2 \quad ; \quad s = K+1, \ldots, n \tag{۶-۴۱}$$

بدیهی است که به‌ازای $s=K$، مقدار RSS_s برابر صفر است (چرا؟).

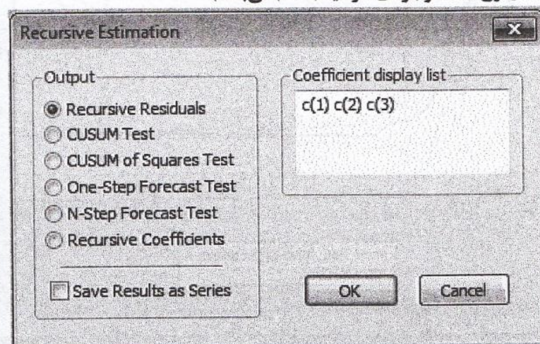
حال اگر ضرایب بازگشتی یا RSSهای بازگشتی را محاسبه و ترسیم کنیم، روند آنها نشان می‌دهد که در چه سالی تغییر ضرایب رخ داده است. معمولاً نرم‌افزارهایی مانند Eviews این محاسبات را انجام داده و نمودار آنها را همراه با «به‌اضافه و منهای دو انحراف معیار» ترسیم می‌کند. اگر RSSها از این مرزها عبور کنند نشان‌دهندهٔ تغییرات ساختاری است.

برآوردهای بازگشتی در Eviews — فایل data2

ابتدا معادله مورد نظر را برآورد می‌کنیم و در پنجرهٔ نتایج، مسیر زیر را دنبال می‌کنیم:

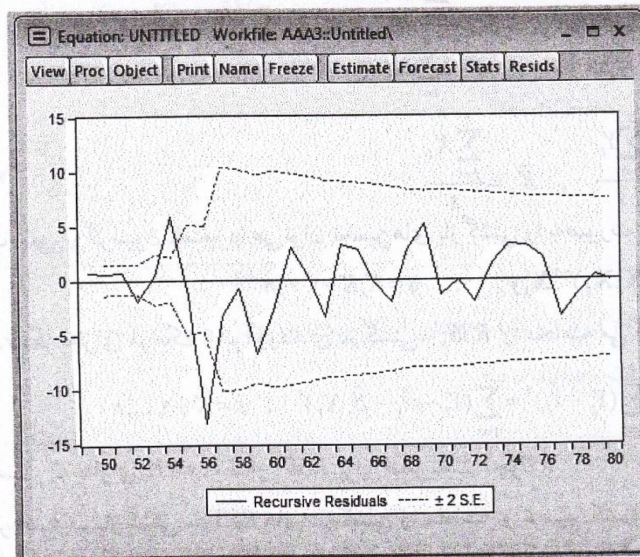
View → Stability Diagnostics → Recursive Estimates (OLS only)

پنجرهٔ زیر باز می‌شود (معادلهٔ برآوردی بیانگر رگرسیون نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) روی عرض از مبدأ، نرخ رشد نیروی کار (GL) و نرخ رشد موجودی سرمایه (GK) می‌باشد):



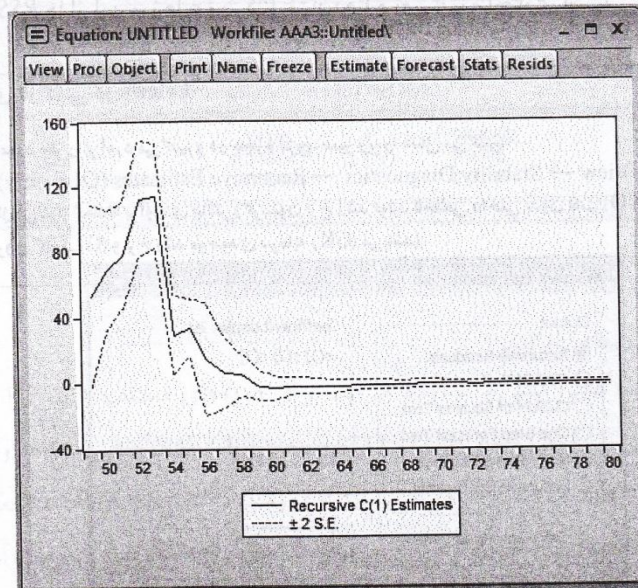
در پنجرهٔ فوق، گزینهٔ Recursive Residuals نمودار و مرز RSSهای بازگشتی را رسم می‌کند. همچنین، گزینه آخر، یعنی Recursive Coefficients، نمودار و مرز هر یک از ضرایب را رسم می‌کند. در سمت راست پنجره فوق، می‌توان هر

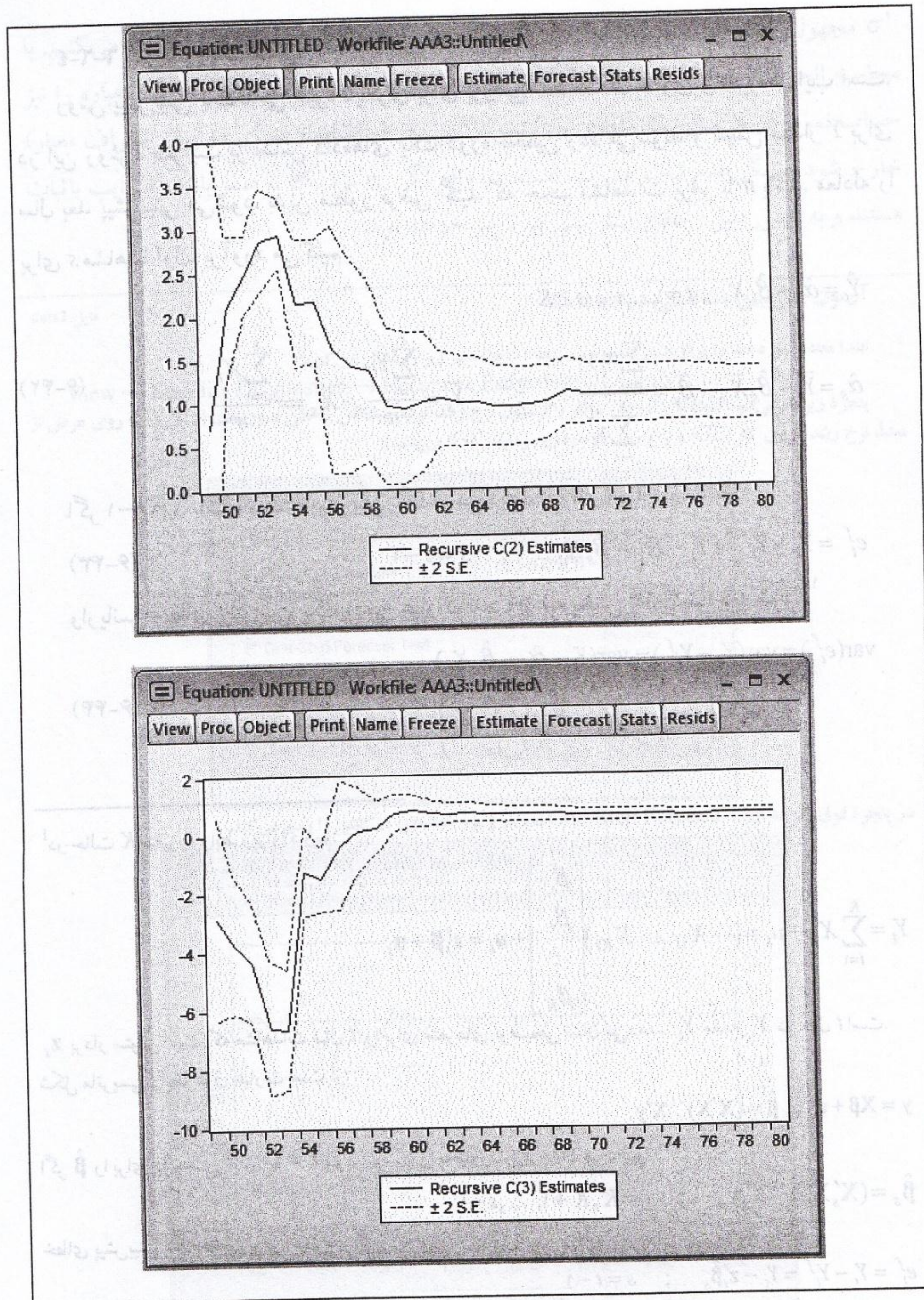
یک از ضرایب را انتخاب نموده و نمودارهای مربوطه را ترسیم نمود. به‌عنوان نمونه، RSSهای بازگشتی برای همه ضرایب عبارت است از:



در سال ۱۳۵۲، ۱۳۵۴ و ۱۳۵۶ تغییر ساختاری رخ داده است، زیرا RSSها از مرز خارج شده‌اند.

برای هر یک از ضرایب، نمودارهای زیر به‌دست آمده است و نشان می‌دهد که ضرایب C(1)، C(2) و C(3) از سال ۱۳۵۸ به بعد تقریباً باثبات بوده‌اند.




Equation: UNTITLED Workfile: AAA3::Untitled\
View Proc Object Print Name Freeze Estimate Forecast Stats Resids
Recursive C(2) Estimates
± 2 S.E.
Equation: UNTITLED Workfile: AAA3::Untitled\
View Proc Object Print Name Freeze Estimate Forecast Stats Resids
Recursive C(3) Estimates
± 2 S.E.

٤-٦-٦ پیش‌بینی یک‌قدمی

روش پیش‌بینی یک‌قدمی شیوه دیگری برای بررسی شکست ساختاری و تغییر ضرایب است. در این روش، ضرایب بر اساس داده‌های یک دوره تخمین زده می‌شوند و سپس مقدار Y برای سال بعد پیش‌بینی می‌شود. بدین منظور فرض کنید که حجم مشاهدات برابر با n باشد. معادله را برای s مشاهدهٔ اول، برآورد می‌کنیم:

$$\hat{Y}_t = \hat{\alpha}_s + \hat{\beta}_s X_t \; ; \; t=۱,\dots,s \; , \; s \geq K$$

$$\hat{\alpha}_s = \overline{Y}_s - \hat{\beta}_s \overline{X}_s \, , \; \hat{\beta}_s = \frac{\sum_{t=۱}^{s} x_t y_t}{\sum_{t=۱}^{s} x_t^۲} \quad ; \quad \overline{Y}_s = \frac{\sum_{t=۱}^{s} Y_t}{n} \, , \; \overline{X}_s = \frac{\sum_{t=۱}^{s} X_t}{n} \qquad (۶-۴۲)$$

اگر $s = t - ۱$ باشد، خطای پیش‌بینی یک‌قدمی (یک‌دوره‌ای) برابر است با:

$$e_t^f = Y_t - Y_t^f = Y_t - \hat{\alpha}_s - \hat{\beta}_s X_t \quad ; \quad s = t - ۱ \qquad (۶-۴۳)$$

واریانس خطای پیش‌بینی یک قدمی عبارت است از (به بخش ۱۴-۲ مراجعه شود):[1]

$$\operatorname{var}(e_t^f) = \operatorname{var}(Y_t - Y_t^f) = \operatorname{var}(Y_t - \hat{\alpha}_s - \hat{\beta}_s X_t)$$
$$= \sigma^۲ \left(۱ + \frac{۱}{n} + (X_t - \overline{X}_s)^۲ \Big/ \sum_{t=۱}^{s} x_t^۲ \right) \quad ; \quad s = t - ۱ \qquad (۶-۴۴)$$

---

[1] در حالت K متغیره روابط زیر را داریم:

$$Y_t = \sum_{i=۱}^{K} X_{it} + u_t = (۱ \;\; X_{۲t} \; \dots \; X_{Kt}) \begin{bmatrix} \beta_۱ \\ \beta_۲ \\ \vdots \\ \beta_K \end{bmatrix} + u_t = z_t' \beta + u_t$$

$\mathbf{z}_t$ بردار ستونی است که مشاهدات سال t را برای متغرهای توضیحی نشان می‌دهد. Y_t مقدار Y در سال t است. شکل ماتریسی رابطه فوق عبارت است از:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \; , \; \hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-۱}\mathbf{X'y}$$

اگر $\hat{\boldsymbol{\beta}}$ را برای داده‌های $t = ۱, \dots s$ بنویسیم، آن را به‌صورت زیر نشان می‌دهیم:

$$\hat{\boldsymbol{\beta}}_s = (\mathbf{X}_s'\mathbf{X}_s)^{-۱}\mathbf{X}_s'\mathbf{y}_s \quad ; \quad s = K, K+۱, \dots, t-۱$$

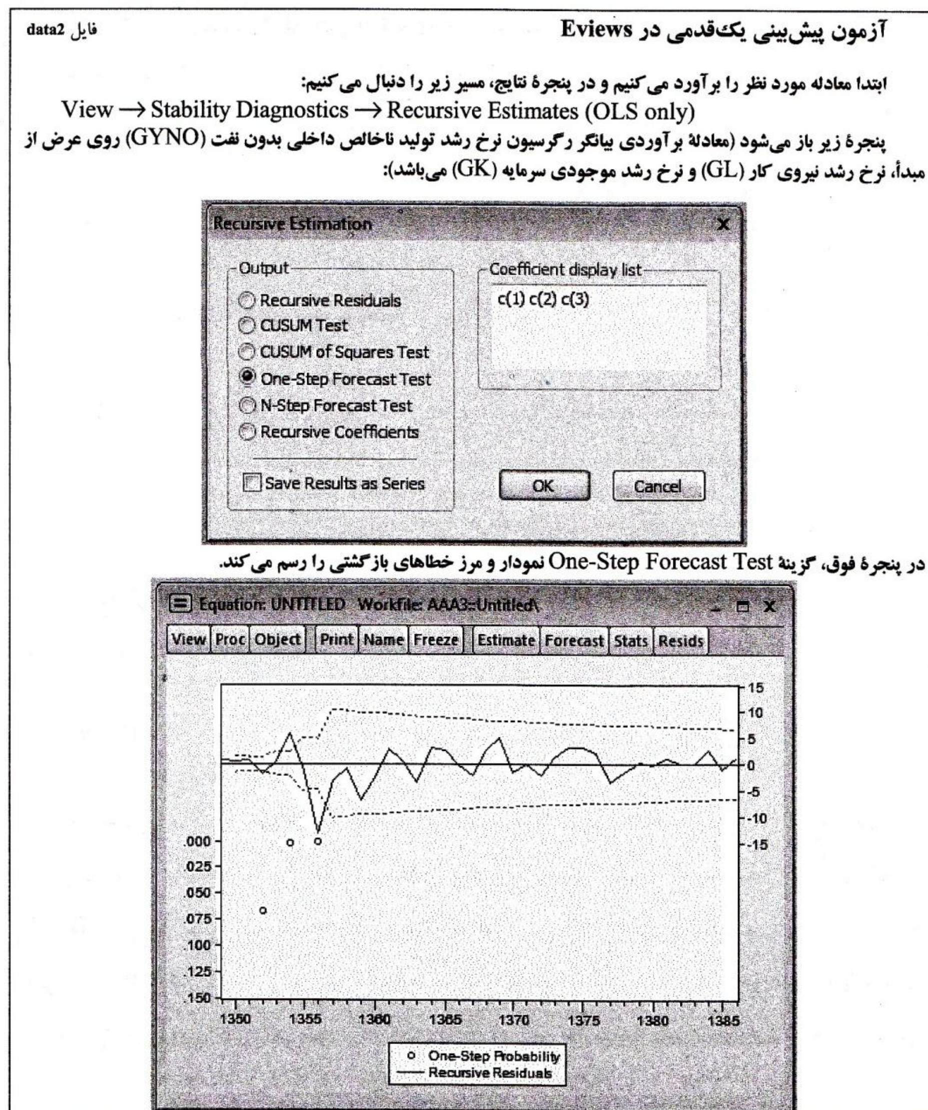
خطای پیش‌بینی یک قدمی برابر است با:

$$e_t^f = Y_t - Y_t^f = Y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_s \quad ; \quad s = t - ۱$$

واریانس خطای پیش‌بینی یک قدمی عبارت است:

$$\operatorname{var}(e_t^f) = \operatorname{var}(Y_t - Y_t^f) = \sigma^۲\left(۱ + \mathbf{z}_t'(\mathbf{X}_s'\mathbf{X}_s)^{-۱}\mathbf{z}_t\right)$$

σ^2 مجهول است که به جای آن از واریانس جملات خطاهای s مشاهدهٔ اول استفاده می‌کنیم. با محاسبه انحراف معیار خطای پیش‌بینی، می‌توان «به‌اضافه و منهای دو برابر انحراف معیار» را نیز حساب نمود. در هر سالی که خطای پیش‌بینی از این مرز (به‌اضافه و منهای دو برابر انحراف معیار) خارج شود، بیانگر عدم ثبات ضرایب است. در واقع اگر e_t^f نزدیک به صفر باشد، ضرایب باثبات هستند و به همین دلیل، خطای پیش‌بینی نیز ناچیز خواهد بود.



آزمون پیش‌بینی یک‌قدمی در Eviews — فایل data2

ابتدا معادله مورد نظر را برآورد می‌کنیم و در پنجرهٔ نتایج، مسیر زیر را دنبال می‌کنیم:

View → Stability Diagnostics → Recursive Estimates (OLS only)

پنجرهٔ زیر باز می‌شود (معادلهٔ برآوردی بیانگر رگرسیون نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) روی عرض از مبدأ، نرخ رشد نیروی کار (GL) و نرخ رشد موجودی سرمایه (GK) می‌باشد):

در پنجرهٔ فوق، گزینهٔ One-Step Forecast Test نمودار و مرز خطاهای بازگشتی را رسم می‌کند.

نمودار فوق نشان می‌دهد که در برخی از سالها (در اینجا سه سال)، خطاهای پیش‌بینی یک‌قدمی، از مرزها فراتر رفته است. محور عمودی سمت راست، مقدار خطاها و مرزها را نشان می‌دهد. محور عمودی سمت چپ، مقدار احتمال نقاطی که از مرز خارج شده‌اند را نشان می‌دهند. مثلاً در سال ۱۳۵٦ که خارج از مرز است، احتمال آن ۰/۰۰۱ است (اگر نشانگر ماوس را روی نقطه موردنظر نگه دارید، سال و مقدار احتمال را مشخص می‌کند). همچنین در سال ۱۳۵۲ احتمال برابر با ۰/۰۰۲۵ و در سال ۱۳۵۲ احتمال برابر با ۰/۰٦۷۱ است. بنابراین، تغییر ضرایب در سال ۱۳۵٤ و ۱۳۵٦ در سطح ۰/۰۵ و حتی در سطح ۰/۰۱ معنادار است، اما در سال ۱۳۵۲ در سطح ۰/۰٦۷ درصد.

٥-٦-٦ آزمون مجموع تجمعی خطاهای بازگشتی (CUSUM)

در مبحث پیش‌بینی یک قدمی برای سال t دیدیم که خطای پیش‌بینی و واریانس آن عبارت است از:

$$e_t^f = Y_t - Y_t^f = Y_t - \hat{\alpha}_s - \hat{\beta}_s X_t \quad ; \quad s = t - ۱$$
$$\operatorname{var}(e_t^f) = \sigma^۲ \left(۱ + \frac{۱}{n} + (X_t - \overline{X}_s)^۲ \Big/ \sum_{t=۱}^{s} x_t^۲ \right) \quad ; \quad s = t - ۱ \qquad (۶\text{-}۴۵)$$

حال خطاهای بازگشتی را به‌صورت زیر می‌نویسیم:

$$w_t = \frac{e_t^f}{\sqrt{۱ + \frac{۱}{n} + \frac{(X_t - \overline{X}_s)^۲}{\sum_{t=۱}^{s} x_t^۲}}} \quad ; \quad s = t - ۱ \;,\; t = K + ۱, \ldots, n \qquad (۶\text{-}۴۶)$$

w_t توزیع نرمال با میانگین صفر و واریانس $\sigma^۲$ دارد. حال w_tها را استاندارد کرده و مقدار تجمعی آنها را حساب می‌کنیم:

$$W_t = \sum_{j=K+۱}^{t} \frac{w_j}{\hat{\sigma}} \quad ; \quad t = K + ۱, \ldots, n \;,\; \hat{\sigma}^۲ = \frac{RSS_n}{n - K} \qquad (۶\text{-}۴۷)$$

RSS_n مجموع مجذور خطاها برای معادله‌ای است که با تمامی مشاهدات برازش شده است. اگر ضرایب باثبات باشند، در این صورت خطای پیش‌بینی ناچیز است و لذا مجموع تجمعی آنها نیز کوچک خواهد بود و در نتیجه، W_t نزدیک به صفر می‌باشد (یعنی نزدیک به امید ریاضی آن، زیرا $E(W_t) = ۰$ است). در غیر این صورت، W_t تفاوت معناداری با صفر خواهد داشت. برای آزمون W_t مرزهایی تعیین می‌شود که اگر W_t از این مرزها عبور کند نشان‌دهندهٔ تغییر ضرایب

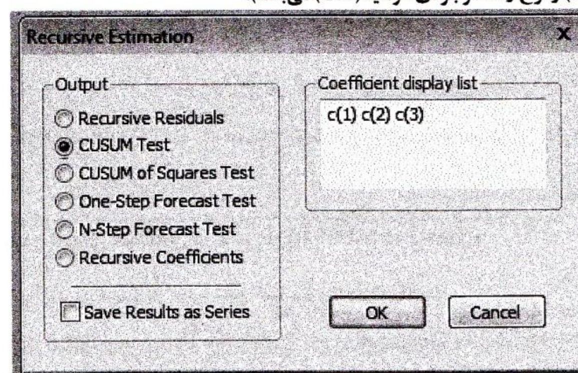
می‌باشد. مختصات شروع این مرزها برابر با $(K,\pm a\sqrt{n-K})$ و مختصات نقطهٔ پایان آن برابر با $(n,\pm ۳a\sqrt{n-K})$ می‌باشد.

آزمون مجموع تجمعی خطاها (CUSUM) در Eviews فایل data2

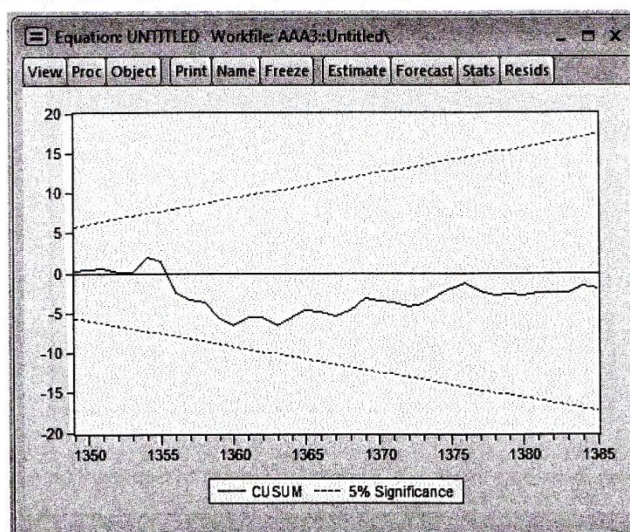
ابتدا معادله مورد نظر را برآورد می‌کنیم و در پنجرهٔ نتایج، مسیر زیر را دنبال می‌کنیم:

View → Stability Diagnostics → Recursive Estimates (OLS only)

پنجرهٔ زیر باز می‌شود(معادلهٔ برآوردی بیانگر رگرسیون نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) روی عرض از مبدأ، نرخ رشد نیروی کار (GL) و نرخ رشد موجودی سرمایه (GK) می‌باشد):



در پنجرهٔ فوق، گزینهٔ CUSUM Test را انتخاب می‌کنیم. نمودار و مرز خطاهای تجمعی بازگشتی به‌صورت زیر ترسیم می‌شود.



نمودار فوق نشان می‌دهد که هیچ یک از خطاهای تجمعی از مرزها خارج نشده‌اند و لذا این آزمون هیچ‌گونه تغییر ساختاری را نشان نمی‌دهد.

۶-۶-۶ آزمون مجموع مجذور تجمعی خطاهای بازگشتی (CUSUMQ)

در این روش، از نسبت زیر استفاده می‌شود:

$$S_t = \frac{\sum_{j=K+1}^{t} w_j^2}{\sum_{j=K+1}^{n} w_j^2} \quad ; \quad t = K+1, \ldots, n \qquad (6\text{-}48)$$

هر یک از w_jها تحت فرضیه H_0 (فرضیه ثبات ضرایب) دارای توزیع χ_1^2 هستند. امید ریاضی S_t برابر با $E(S_t) = \frac{t-K}{n-K}$ می‌باشد که بین صفر و ۱ است. اگر انحراف از میانگین، زیاد باشد در این صورت، تغییر ضرایب رخ داده است. مرزها نیز توسط نرم‌افزارهایی مانند Eviews ترسیم می‌شود. مرزها به‌صورت $E(S_t) \pm C$ می‌باشد که مقدار C بستگی به سطح معناداری دارد.

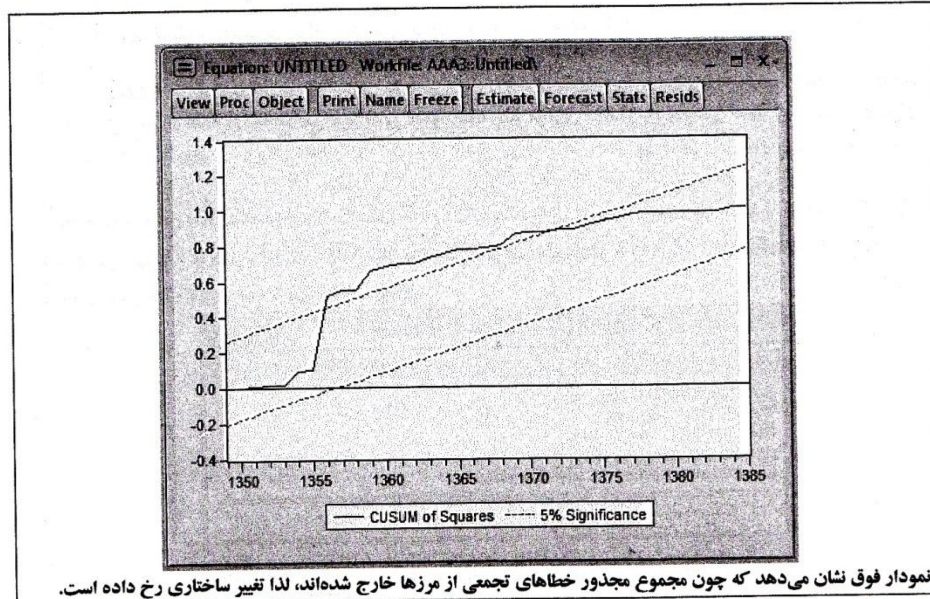
آزمون مجموع مجذور تجمعی خطاهای بازگشتی (CUSUMQ) در Eviews — فایل data2

ابتدا معادله مورد نظر را برآورد می‌کنیم و در پنجرهٔ نتایج، مسیر زیر را دنبال می‌کنیم:

View → Stability Diagnostics → Recursive Estimates (OLS only)

پنجرهٔ زیر باز می‌شود (معادلهٔ برآوردی بیانگر رگرسیون نرخ رشد تولید ناخالص داخلی بدون نفت (GYNO) روی عرض از مبدأ، نرخ رشد نیروی کار (GL) و نرخ رشد موجودی سرمایه (GK) می‌باشد):

در پنجرهٔ فوق، گزینهٔ CUSUM of Squares Test را انتخاب می‌کنیم. نمودار و مرز مجذور خطاهای تجمعی بازگشتی به‌صورت زیر ترسیم می‌شود.



نمودار فوق نشان می‌دهد که چون مجموع مجذور خطاهای تجمعی از مرزها خارج شده‌اند، لذا تغییر ساختاری رخ داده است.

۷-۶-۶ متغیرهای مجازی[1] و آزمون ثبات ضرایب

فرضیه ثبات پارامترها را می‌توان با استفاده از متغیرهای مجازی نیز انجام داد. بدین منظور فرض کنید که مدل (۳۰-۶) را داشته باشیم. کل دوره مورد مطالعه را به دو دوره تقسیم می‌کنیم که دوره اول شامل $n_۱$ مشاهده و در دوره دوم شامل $n_۲$ مشاهده می‌باشد ($n_۱ + n_۲ = n$). مدل (۳۰-۶) بیانگر رگرسیون مقید است که برای کل دوره برآورد می‌شود. اما رگرسیون غیرمقید را برای کل دوره به‌صورت زیر تعریف می‌کنیم:

$$Y_t = \beta_۱ + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + \beta_۴ D_t + \beta_۵ D_t X_{۲t} + \beta_۶ D_t X_{۳t} + u_t \qquad (۶\text{-}۴۹)$$

D_t متغیر مجازی است که مقدار آن در دوره اول برابر با ۱ و در دوره دوم برابر با صفر است. اگر قید $\beta_۴ = \beta_۵ = \beta_۶ = ۰$ را اعمال کنیم، معادله (۴۹-۶) تبدیل به (۳۰-۶) خواهد شد. اگر ضرایب در طی دوره اول و دوم ثابت باشند، در این صورت ضرایب $\beta_۴$، $\beta_۵$ و $\beta_۶$ از نظر آماری معنادار نخواهند بود.

دوره اول: $D_t = ۱ \Rightarrow Y_t = (\beta_۱ + \beta_۴) + (\beta_۲ + \beta_۵) X_{۲t} + (\beta_۳ + \beta_۶) X_{۳t} + u_t$

دوره دوم: $D_t = ۰ \Rightarrow Y_t = \beta_۱ + \beta_۲ X_{۲t} + \beta_۳ X_{۳t} + u_t$

[1] جزئیات بیشتر در فصل پنجم ارائه شده است.

آزمون ثبات ضرایب با استفاده از متغیرهای مجازی در Eviews — فایل data2

آزمون ثبات ضرایب تابع سرمایه‌گذاری خصوصی (IP) را با استفاده از متغیرهای مجازی انجام می‌دهیم. در اینجا دو دوره قبل و بعد از انقلاب را مقایسه می‌کنیم. بدین منظور ابتدا متغیر مجازی D57 را تعریف می‌کنیم که مقدار آن برای سالهای ۵۶-۱۳۴۶ برابر با ۱ و برای سالهای ۸۰-۱۳۵۷ برابر با صفر می‌باشد. حال مدل مورد نظر را با دستور زیر برآورد می‌کنیم:

```
LS    IP  C  Y  R  D57   D57*Y   D57*R
```

توجه شود که این مدل مشابه مدل (۳۲-۶) است که در آن C مقدار ثابت است که برابر با β_1 می‌باشد. همچنین β_2 و β_3 بیانگر ضرایب Y و R هستند. اما β_4، β_5 و β_6 به ترتیب ضرایب متغیرهای D57، D57*Y و D57*R می‌باشد. با برآورد مدل فوق نتایج به‌صورت زیر به‌دست می‌آید:

Equation: UNTITLED Workfile: YNO::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: IP
Method: Least Squares
Date: 01/30/11 Time: 18:48
Sample (adjusted): 1345 1380
Included observations: 36 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-28338.87	12621.01	-2.245372	0.0323
Y	0.393696	0.109819	3.584941	0.0012
R	-2833.283	1208.771	-2.343937	0.0259
D57	22437.89	27178.52	0.825574	0.4156
D57*Y	-0.147661	0.153192	-0.963898	0.3428
D57*R	1105.481	4931.474	0.224168	0.8241

R-squared	0.665818	Mean dependent var	25513.35
Adjusted R-squared	0.610121	S.D. dependent var	13257.30
S.E. of regression	8277.896	Akaike info criterion	21.03158
Sum squared resid	2.06E+09	Schwarz criterion	21.29550
Log likelihood	-372.5684	Hannan-Quinn criter.	21.12369
F-statistic	11.95430	Durbin-Watson stat	1.447742
Prob(F-statistic)	0.000002		

ملاحظه می‌شود که هیچ یک از ضرایب مربوط به متغیرهای مجازی، معنی‌دار نیستند و لذا ثبات ضرایب رد نمی‌شود. بدین ترتیب نتایج حاصل از روش چاو و روش متغیرهای مجازی یکسان می‌باشد.

۷-۶ خطای اندازه‌گیری متغیرها

یکی از مشکلاتی که در عمل با آن مواجه می‌شویم، خطای اندازه‌گیری است. به هر دلیلی ممکن است داده‌های گردآوری شده دارای خطا باشند. این خطا می‌تواند هم در مورد متغیر وابسته و هم متغیرهای توضیحی باشد.

ابتدا این بحث را برای متغیرهای توضیحی بررسی می‌کنیم. فرض کنید که مقادیر واقعی متغیر توضیحی برابر با X^* باشد. در این صورت مدل زیر را داریم:

$$Y_t = \alpha + \beta X_t^* + u_t \qquad (۶-۵۰)$$

اما مقدار واقعی X (یعنی X^*) قابل دسترس نمی‌باشد و لذا از X استفاده می‌کنیم. فرض کنید که بین X و X^* رابطه زیر برقرار باشد:

$$X_t = X_t^* + \varepsilon_t \tag{۶-۵۱}$$

بنابراین X_t دارای خطا است که آن را با ε_t نشان می‌دهیم. اما فرض می‌شود که ε_t یک متغیر تصادفی با میانگین صفر و واریانس $\sigma_\varepsilon^۲$ است. با جایگذاری (۶-۵۱) در (۶-۵۰) خواهیم داشت:

$$Y_t = \alpha + \beta X_t + (u_t - \beta\varepsilon_t) \tag{۶-۵۲}$$

$$= \alpha + \beta X_t + v_t \quad , \quad v_t = u_t - \beta\varepsilon_t$$

در اینجا جمله خطا v_t است که مستقل از متغیر توضیحی نیست، زیرا:

$$\text{cov}(v_t , X_t) = E[(v_t - Ev_t)(X_t - EX_t)] = E(v_t\varepsilon_t) = \text{cov}(v_t, \varepsilon_t)$$

زیرا $E(\varepsilon_t) = E(v_t) = ۰$ و $E(X_t) = E(X_t^*) + E(\varepsilon_t) = X_t^*$ است. حال به‌جای v_t قرار داده و بقیه محاسبات را انجام می‌دهیم:

$$\text{cov}(X_t, v_t) = E(v_t\varepsilon_t) = E((u_t - \beta\varepsilon_t)\varepsilon_t)$$

$$= E(u_t\,\varepsilon_t) - \beta E(\varepsilon_t^۲) = ۰ - \beta\sigma_\varepsilon^۲ = -\beta\sigma_\varepsilon^۲ \tag{۶-۵۳}$$

$E(u_t\,\varepsilon_t) = ۰$ است، زیرا u_t و ε_t مستقل‌اند.

حال تخمین β را از مدل (۶-۵۲) به‌صورت زیر می‌نویسیم و از معادله (۶-۵۰) که مقدار واقعی Y_t را توصیف می‌کند قرار می‌دهیم:

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^۲} = \frac{\sum x_t Y_t}{\sum x_t^۲} = \frac{\sum x_t(\alpha + \beta X_t^* + u_t)}{\sum x_t^۲}$$

$$\hat{\beta} = \frac{\beta\sum x_t X_t^* + \sum x_t u_t}{\sum x_t^۲} = \frac{\beta\sum x_t x_t^* + \sum x_t(u_t - \bar{u})}{\sum x_t^۲}$$

به‌جای x_t از (۶-۵۱) مقدار $x_t^* + (\varepsilon_t - \bar{\varepsilon})$ را قرار می‌دهیم:

$$\hat{\beta} = \frac{\beta\sum\left(x_t^* + (\varepsilon_t - \bar{\varepsilon})\right)x_t^* + \sum\left(x_t^* + (\varepsilon_t - \bar{\varepsilon})\right)(u_t - \bar{u})}{\sum\left(x_t^* + (\varepsilon_t - \bar{\varepsilon})\right)^۲}$$

با ساده نمودن عبارت فوق و محاسبه حد احتمال (plim) آن خواهیم داشت:

$$\text{plim}\,\hat{\beta}=\frac{\beta\sigma_{x^*}^{2}}{\sigma_{x^*}^{2}+\sigma_{\varepsilon}^{2}}=\beta\frac{1}{1+\frac{\sigma_{\varepsilon}^{2}}{\sigma_{x^*}^{2}}}<\beta \qquad (۶-۵۴)$$

که $\text{plim}\frac{\Sigma x_t^{*2}}{n}=\sigma_{x^*}^{2}$ و $\text{plim}\frac{\Sigma(\varepsilon_t-\bar{\varepsilon})^{2}}{n}=\sigma_{x^*}^{2}$ و برای سایر جملات برابر با صفر می‌باشد، زیرا $\text{cov}(x^*,\varepsilon)=\text{cov}(x^*,u)=\text{cov}(u,\varepsilon)=0$ است. بنابراین حتی اگر حجم نمونه نیز افزایش یابد، تورش $\hat{\beta}$ از بین نخواهد رفت. با توجه به اینکه $\frac{1}{1+\sigma_{\varepsilon}^{2}/\sigma_{x^*}^{2}}$ کوچکتر از ۱ است، لذا $\hat{\beta}$ کمتر از حد واقعی، برآورد می‌گردد.

اکنون حالتی را بررسی می‌کنیم که خطای اندازه‌گیری مربوط بـه متغیـر وابـسته باشـد. در ایـن صورت مدل واقعی به‌صورت زیر می‌باشد:

$$Y_t^*=\alpha+\beta X_t+u_t \qquad (۶-۵۵)$$

Y_t^* مقادیر واقعی Y است، اما به‌جای آن، Y_t را استفاده می‌کنیم که دارای خطا می‌باشد:

$$Y_t=Y_t^*+z_t \qquad (۶-۵۶)$$

که z_t خطای اندازه‌گیری را نشان مـی‌دهـد. حـال در معادلـه (۳۸-۶) بـه‌جـای Y_t^* از (۵۶-۶) قـرار می‌دهیم:

$$Y_t-z_t=\alpha+\beta X_t+u_t \quad\Rightarrow\quad Y_t=\alpha+\beta X_t+(u_t+z_t)$$

$$Y_t=\alpha+\beta X_t+w_t \quad , \quad w_t=u_t+z_t \qquad (۶-۵۷)$$

تخمین β برای مدل (۵۷-۶) عبارت است از:

$$\hat{\beta}=\frac{\sum x_t y_t}{\sum x_t^{2}}=\frac{\sum x_t Y_t}{\sum x_t^{2}}$$

به‌جای Y_t از مدل (۵۷-۶) قرار می‌دهیم:

$$\hat{\beta}=\frac{\sum x_t[\alpha+\beta X_t+w_t]}{\sum x_t^{2}}=\beta+\frac{\sum x_t w_t}{\sum x_t^{2}} \quad\Rightarrow E(\hat{\beta})=\beta$$

بنابراین، تخمین‌زنندهٔ OLS همچنان بدون تورش است.

همان‌طور که در رگرسیون یک متغیره ثابت شد، واریانس $\hat{\beta}$ برابر است با:

$$\text{var}(\hat{\beta})=\frac{\sigma_w^{2}}{\sum x_t^{2}} \qquad (۶-۵۸)$$

اما با توجه به $w_t = u_t + z_t$ و با توجه به استقلال z_t و u_t، رابطه $\sigma_w^2 = \sigma_u^2 + \sigma_z^2$ برقرار است:

$$\text{var}(\hat{\beta}) = \frac{\sigma_u^2 + \sigma_z^2}{\sum x_t^2} \qquad (۶-۵۹)$$

لذا واضح است که خطای اندازه‌گیری در متغیر وابسته باعث افزایش واریانس تخمین‌زنندهٔ OLS می‌شود، اما نمی‌تواند موجب تورش شود، زیرا $E(\hat{\beta}) = \beta$ می‌باشد.

۶-۹ علّیت

علّیت یکی از مسائل اساسی در بررسی رابطه بین متغیرهای اقتصادی است. تعیین جهت علّیت برای متغیرهایی مورد استفاده قرار می‌گیرد که مبانی نظری صریحی در مورد آنها وجود ندارد. یکی از مثال‌های معروف در اقتصاد که مورد مجادله می‌باشد مربوط به رابطهٔ بین رشد تولید ناخالص ملی (Y) و رشد پول (X) است. سؤال این است که آیا رشد پول موجب رشد تولید ناخالص ملی می‌شود یا اینکه ابتدا تولید ناخالص ملی افزایش می‌یابد و سپس موجب افزایش نیاز به پول می‌گردد و به‌دنبال آن بانک مرکزی حجم پول را افزایش می‌دهد.

روش مرسوم برای بررسی علّیت، معروف به آزمون علّیت گرانجر[1] است. در این روش معادلات زیر مورد بررسی قرار می‌گیرند:

$$Y_t = \sum_{i=1}^{n} \alpha_i X_{t-i} + \sum_{j=1}^{n} \beta_j Y_{t-j} + u_t \qquad (۶-۸۳)$$

$$X_t = \sum_{i=1}^{n} a_i x_{t-i} + \sum_{j=1}^{n} b_j Y_{t-j} + v_t \qquad (۶-۸۴)$$

بر اساس معادلات فوق می‌توان بدین صورت بحث نمود که:

الف) اگر $\sum \alpha_i \neq 0$ و $\sum b_j = 0$ بوده و از نظر آماری معنی‌دار باشند، در این صورت علّیت از X به Y است.

ب) اگر $\sum \alpha_i = 0$ و $\sum b_j \neq 0$ باشد در این صورت علّیت از Y به X است.

ج) اگر $\sum \alpha_i \neq 0$ و $\sum b_j \neq 0$ باشد، در این صورت علّیت دو طرفه است.

---

[1] Granger

د) اگر $\sum \alpha_i = 0$ و $\sum b_j = 0$ باشد، این دو متغیر مستقل‌اند و رابطه‌ای با هم ندارند.

هر یک از این فرضیه‌ها را می‌توان با آماره F که قبلاً برای مقایسه رگرسیون‌های مقید و غیرمقید معرفی گردید، آزمون نمود. هر یک از فرضیه‌های فوق بیانگر اعمال یک قید روی ضرایب معادله موردنظر است که با آماره F قابل آزمون است.

آزمون علّیت گرانجر در Eviwes فایل data5

برای آزمون علّیت در Eviews به‌صورت زیر عمل می‌کنیم:

Quick → Group Statistics → Granger Causality Test

با انتخاب مسیر فوق، پنجره‌ای با عنوان Series List باز می‌شود که در آن نام متغیرهای مورد نظر را وارد می‌کنیم.

Series List

List of series, groups, and/or series expressions

y1 x1

OK Cancel

با انتخاب OK پنجره کوچکی با عنوان Lag Specification باز می‌شود که در آن تعداد وقفه‌هایی را که لازم است به متغیرها بدهیم وارد می‌کنیم.

Lag Specification

Lags to include: 9

OK Cancel

در اینجا لازم است که وقفه‌های مختلف را بررسی کنیم، زیرا ممکن است بین دو متغیر، در وقفه‌های کوتاه رابطهٔ علیت وجود نداشته باشد، ولی با وقفه‌های طولانی‌تر بین آنها رابطهٔ علیت به‌وجود آید. به‌عنوان مثال نتایج آزمون علیت برای دو متغیر X و Y به‌صورت زیر داده می‌شود:

د) اگر $\sum \alpha_i = ۰$ و $\sum b_j = ۰$ باشد، این دو متغیر مستقل‌اند و رابطه‌ای با هم ندارند.

هر یک از این فرضیه‌ها را می‌توان با آماره F که قبلاً برای مقایسه رگرسیون‌های مقید و غیرمقید معرفی گردید، آزمون نمود. هر یک از فرضیه‌های فوق بیانگر اعمال یک قید روی ضرایب معادله موردنظر است که با آماره F قابل آزمون است.

آزمون علّیت گرانجر در Eviwes فایل data5

برای آزمون علّیت در Eviews به‌صورت زیر عمل می‌کنیم:

Quick → Group Statistics → Granger Causality Test

با انتخاب مسیر فوق، پنجره‌ای با عنوان Series List باز می‌شود که در آن نام متغیرهای مورد نظر را وارد می‌کنیم.

Series List

List of series, groups, and/or series expressions

y1 x1

OK Cancel

با انتخاب OK پنجره کوچکی با عنوان Lag Specification باز می‌شود که در آن تعداد وقفه‌هایی را که لازم است به متغیرها بدهیم وارد می‌کنیم.

Lag Specification

Lags to include: 9

OK Cancel

در اینجا لازم است که وقفه‌های مختلف را بررسی کنیم، زیرا ممکن است بین دو متغیر، در وقفه‌های کوتاه رابطهٔ علیت وجود نداشته باشد، ولی با وقفه‌های طولانی‌تر بین آنها رابطهٔ علیت به‌وجود آید. به‌عنوان مثال نتایج آزمون علیت برای دو متغیر X و Y به‌صورت زیر داده می‌شود:

Group: UNTITLED Workfile: DATA5::Untitled\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Pairwise Granger Causality Tests
Date: 12/26/13 Time: 14:20
Sample: 1353 1386
Lags: 9

Null Hypothesis:	Obs	F-Statistic	Prob.
X1 does not Granger Cause Y1	23	6.64369	0.0419
Y1 does not Granger Cause X1		40.5856	0.0014

در جدول فوق، عبارت «X does not Granger Cause Y» بیانگر این فرضیه است که «X علت Y نیست». این فرضیه معادل با این است که در معادلات (۶-۶۴) و (۶-۶۵)، $\sum \alpha_i = ۰$ و $\sum b_j \neq ۰$ می‌باشد. حال برای این فرضیه، اگر مقدار F بزرگتر از F جدول باشد، یا مقدار احتمال آن که در ستون آخر داده می‌شود، کوچکتر از ۰/۰۵ باشد، در این صورت فرضیه فوق رد می‌شود. این بدان معنا است که «X علت Y است». بنابراین در مثال فوق، با توجه به مقادیر F و احتمال‌های آن، «X علت Y است» و همچنین «Y علت X است». پس رابطه علیت دو طرفه است.

مسائل

۱- محاسبات زیر را داریم:

$$\mathbf{X'X} = \begin{bmatrix} ۲۰ & ۱۰ \\ ۱۰ & ۳۰ \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} ۳۰ \\ ۴۰ \end{bmatrix} \qquad \sum_i y_i^۲ = ۷۵$$

الف) رگرسیون Y روی X را برآورد کنید.

ب) یک مشاهده جدید با $X = ۲$ و $Y = ۴$ به‌دست آمده است، آزمون پیش‌بینی چاو را برای ثبات ضرایب انجام دهید.

۲- یک رگرسیون سه متغیره با استفاده از داده‌های دوره ۸۶-۱۳۶۸ برآورد شده است:

$$\hat{Y}_t = ۲/۲ + ۰/۱۰۴X_{۲t} - ۳/۴۸X_{۳t} + ۰/۳۴X_{۴t}$$

مجموع تغییرات توضیح داده شده ۱۰۹/۶ و مجموع تغییرات توضیح داده نشده ۱۸/۴۸ بود. وقتی رگرسیون را با اضافه نمودن سه متغیر مجازی فصلی برآورد نمودیم مجموع تغییرات توضیح داده شده به ۱۱۴/۸ افزایش یافت. آیا اثرات فصلی معنادار است.

۳- معادله رگرسیون $Y_t = \alpha + \beta X_t + u_t$ با یک نمونه ۲۰تایی از یک منطقه شهری و یک نمونه ۱۰تایی از یک منطقه روستایی در دست بررسی است. مقادیر زیر برای این مناطق حساب شده است:

$$\mathbf{X'X} = \begin{bmatrix} ۲۰ & ۲۰ \\ ۲۰ & ۲۵ \end{bmatrix} \quad \mathbf{X'y} = \begin{bmatrix} ۱۰ \\ ۲۰ \end{bmatrix} \quad \sum y_t^۲ = ۳۰ \quad \text{شهری}$$

$$\mathbf{X'X} = \begin{bmatrix} ۱۰ & ۱۰ \\ ۱۰ & ۲۰ \end{bmatrix} \quad \mathbf{X'y} = \begin{bmatrix} ۸ \\ ۲۰ \end{bmatrix} \quad \sum y_t^۲ = ۲۴ \quad \text{روستایی}$$

آیا برای هر دو منطقه رابطهٔ یکسانی وجود دارد.

۴- در مسئله ۳ موارد زیر را بررسی کنید:

الف) فقط عرض از مبدأ متفاوت باشد.

ب) فقط شیب متفاوت باشد.

۵- در مورد مفهوم همخطی در مدل‌های رگرسیون، به سؤالات زیر پاسخ دهید.

الف) مفهوم همخطی را توضیح دهید.

ب) نحوه شناسایی همخطی را توضیح دهید.

ج) چه راه‌کارهایی را برای رفع همخطی پیشنهاد می‌دهید.

۶- در مدل $Y_t = \beta_۱ + \beta_۲ X_t + \beta_۳ Z_t + u_t$ اگر متغیر Z_t حذف شود، ثابت کنید که موجب تورش $\beta_۲$ می‌شود. همچنین نشان دهید که مقدار تورش بستگی به ضریب همبستگی X و Z دارد.

۷- در مسئله۱ فصل اول، آزمون حذف متغیر Z را انجام دهید.

۸- در مسئله۱ فصل اول، آزمون متغیر زائد را برای Z انجام دهید.

۹- مراحل آزمون فرضیه را برای مقایسه دو رگرسیون $Y_t = \beta_۱ + \beta_۲ X_t + \beta_۳ Z_t + u_t$ و $Y_t = \beta_۱ + \beta_۲ X_t + v_t$ تشریح کنید.

۱۰- در مسئله۱ فصل اول، با استفاده از متغیرهای مجازی بررسی کنید که آیا وقتی مقادیر Z منفی است، اثر آن متفاوت از حالتی است که مقادیر Z مثبت است؟

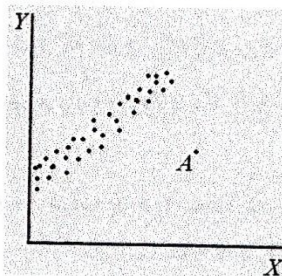
۱۱- در یک رگرسیون پویا میانگین وقفه و میانه وقفه به ترتیب برابر با ۴/۵ و ۳ به‌دست آمده است. مفهوم این مقادیر را به‌طور دقیق توضیح دهید.

۱۲- در مسئله۱ فصل اول، آزمون علیت را برای متغیرهای X، Y و Z انجام دهید.

۱۳- در فرضیه انتظارات تطبیقی و فرضیه تعدیلات جزئی، جملات خطای رگرسیون چه تفاوتی دارند؟

۱۴- در معادله $Y_t = \alpha + \beta X_t + u_t$ اگر هم X و هم Y دارای خطای اندازه‌گیری باشند، چه تأثیری بر تخمین‌زنندهٔ β دارد؟

۱۵- فرض کنید نمودار مشاهدات X و Y به‌صورت زیر باشد.



بعد از برآورد معادله $Y_t = \alpha + \beta X_t + u_t$ معلوم شد که β معنادار نیست. با حذف مشاهده A و برآورد مجدد مدل، β معنادار شد. دلیل آن را بیان کنید.

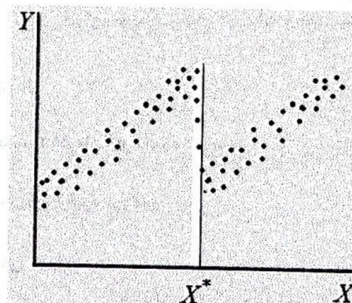
۱۶- ثابت کنید که اگر متغیر Z_t از معادله $Y_t = \alpha + \beta X_t + \gamma Z_t + u_t$ حذف شود و Z_t هیچ همبستگی با X_t نداشته باشد، آنگاه تخمین ضریب X_t در معادله $Y_t = \alpha + \beta X_t + u_t$ بدون تورش خواهد بود.

۱۷- اگر Z_t از معادله $Y_t = \alpha + \beta X_t + \gamma Z z_t + u_t$ حذف شود و Z_t با X_t همبستگی نداشته باشد، آیا تخمین عرضه از مبدأ تغییر خواهد کرد یا نه؟چرا؟

۱۸- درآزمون چاو اگر دونقطه تغییر ساختاری وجود داشته باشد، مراحل انجام آزمون را نوشته و تابع F را تشکیل دهید.

۱۹- ثابت کنید که در معادله $Y_t = \alpha + \beta X_t + u_t$ اگر α را حذف و معادله $Y_t = \beta X_t + u_t$ را برآورد کنیم، آنگاه برآورد β دچار تورش خواهد شد. مقدار تورش به چه چیزی بستگی دارد؟

۲۰- مشاهدات X و Y به‌صورت نمودار زیر است:



الف) با استفاده از متغیرهای مجازی، معادله‌ای را برای توصیف نمودار فوق طرح کنید.

ب) در چه صورتی می‌توان آزمون نقطهٔ شکست چاو را برای نمودار فوق به‌کار برد.

۲۱- معادله $Y_t^* = \alpha + \beta X_t^* + u_t$ را داریم که Y^* و X^* مقادیر واقعی هستند ولی به‌جای آن از $Y_t = \alpha + \beta X_t + u_t$ استفاده می‌شود که $Y_t = Y_t^* + w_t$ و $X_t = Y_t^* + v_t$ می‌باشند. v_t و w_t خطای اندازه‌گیری هستند که میانگین صفر و واریانس σ_v^2 و σ_w^2 دارند. ثابت کنید که تخمین‌زنندهٔ OLS برای β ناسازگاز است.

۲۲- به‌جای مدل $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ معادلات زیر برآوردمی‌شود:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + \varepsilon_t$$
$$X_{2t} = a_1 + a_2 X_{3t} + \varepsilon_t$$

الف) رابطه بین β_1 با α_1 و a_1 چیست؟

ب) رابطه بین β_2 با α_2 و a_2 چیست؟

ج) واریانس $\hat{\beta}_2$ و $\hat{\alpha}_2$ را مقایسه کنید.

۲۳- ثابت کنید که اگر در معادله $Y_t = \alpha + \beta X_t + u_t$ ، متغیر X_t در معرض خطای اندازه‌گیری باشد، موجب تورش در تخمین β خواهدشد.

۲۴- ثابت کنید که اگر در معادله $Y_t = \alpha + \beta X_t + \gamma Z_t + u_t$ به اشتباه Z_t حذف شود و مدل $Y_t = a + bX_t + \varepsilon_t$ برآورد گردد (حذف متغیر مهم) آنگاه تخمین ضریب X_t دارای تورش (اریب) خواهد بود.

ضمیمه فصل ۶: آزمون‌های خاص در Stata

آزمون حذف متغیرهای مهم در Stata

فایل data5

برای انجام این آزمون می‌توان از آزمون F برای مقایسه دو مدل غیر مقید و مقید استفاده نمود. مراحل آزمون عبارت است از:

۱- مدل غیر مقید:

الف) مدل غیر مقید را تخمین می‌زنیم:

```
reg y1  x1  x2
```

ب) مجموع مجذور خطاهای غیر مقید را با فرمان زیر ذخیره می‌کنیم:

```
scalar  rssu=e(rss)
```

ج) درجه آزادی غیر مقید را با فرمان زیر ذخیره می‌کنیم:

```
scalar dfu=e(df_r)
```

۲- مدل مقید (متغیر x2 را حذف می‌کنیم):

الف) مدل مقید را تخمین می‌زنیم:

```
reg y1  x1
```

ب) مجموع مجذور خطاهای مقید را با فرمان زیر ذخیره می‌کنیم:

```
scalar  rssr=e(rss)
```

ج) درجه آزادی مقید را با فرمان زیر ذخیره می‌کنیم:

```
scalar dfr=e(df_r)
```

۳- آماره F را با فرمان زیر حساب می‌کنیم:

```
f=dfu/(dfr-dfu)*(rssr-rssu)/rssu
```

علاه بر این، می‌توان مقدار احتما (pvalue) را با فرمان زیر حساب نمود:

```
scalar pvalue=F(dfr-dfu,dfu,f)
```

٤- مقادیر محاسبه شده را می‌توان با فرمان زیر مشاهده نمود(البته مشاهدهٔ F و pvalue کفایت می‌کند):

```
Scalar list rssu rssr dfu dfr f pvalue
```

مقدار F را با عدد بحرانی مقایسه کرده و نتیجه‌گیری می‌کنیم. اگر مقدار F در ناحیه بحرانی قرار بگیرد نشان می‌دهد که متغیر حذف شده، مهم بوده است. از طرف دیگر اگر مقدار احتمال از ۰/۰۵ کوچکتر باشد، نشان می‌دهد که مقدار F در ناحیه بحرانی قرار دارد و لذا متغیر حذف شده مهم بوده است.

به عنوان مثال نتایج این محاسبات عبارت است از:

```
. scalar list rssr rssu dfr dfu f pvalue
      rssr =  .80476775
      rssu =  .10019204
       dfr =         30
       dfu =         29
         f =  203.93533
    pvalue =  1.185e-14
```

چون مقدار F= ۲۰۳/۹۳ در ناحیه بحرانی قرار دارد (زیرا احتمال کوچکتر از ۰/۰۵ است)، لذا متغیر حذف شدهٔ x2 مهم بوده است.

آزمون متغیرهای زائد

بدین منظور مراحل زیر را انجام می‌دهیم:

۱- معادله مورد نظر را تخمین می‌زنیم:

```
reg  y1 x1 x2
```

۲- فرمان زیر را برای آزمون معنادار بودن x2 انجام می دهیم:

```
test x2=0
```

۳- نتیجه آزمون به صورت زیر به‌دست می‌آید:

```
. test x2=0

 ( 1)  x2 = 0

       F(  1,     29) =  203.94
            Prob > F =    0.0000
```

چون مقدار F= ۲۰۳/۹۳ در ناحیه بحرانی قرار دارد (زیرا احتمال کوچکتر از ۰/۰۵ است)، لذا متغیر x2 زائد نیست.

آزمون چاو در Stata

مراحل آزمون چاو به‌صورت زیر می‌باشد:

۱- ابتدا زمان تغییر ساختاری را معلوم کنید (در این مثال سال ۱۳۷۰ می‌باشد).

۲- یک متغیر مجازی با فرمان زیر ایجاد کنید که برای سال‌های قبل از ۱۳۷۰ برابر با ۱ می‌باشد:

```
gen d=(t>=1370)
```

۳- رگرسیون دوره اول را برآورد کرده و سپس مجموع مجذور خطاها و درجه آزادی را ذخیره کنید:

```
reg y1 x1
scalar rss1=e(rss)
scalar df1=e(df_r)
```

۴- رگرسیون دوره دوم را برآورد کرده و سپس مجموع مجذور خطاها و درجه آزادی را ذخیره کنید:

```
reg y1 x1
scalar rss2=e(rss)
scalar df2=e(df_r)
```

۵- مجموع مجذور خطاهای دو دوره و مجموع درجه آزادی‌ها را حساب کنید:

```
scalar rssu=rss1+rss2
scalar dfu=df1+df2
```

۶- رگرسیون کل دوره را برآورد کرده و سپس مجموع مجذور خطاها و درجه آزادی را ذخیره کنید:

```
reg y1 x1
scalar rssr=e(rss)
scalar dfr=e(df_r)
```

۷- آماره F را و احتمال مربوطه را حساب می‌کنیم:

```
f=dfu/(dfr-dfu)*(rssr-rssu)/rssu
scalar pvalue=F(dfr-dfu, dfu, f)
```

۸- مقادیر محاسبه شده را می‌توان با فرمان زیر مشاهده نمود:

```
scalar list  f  pvalue
```

مقدار F را با عدد بحرانی مقایسه کرده و نتیجه‌گیری می‌کنیم. اگر مقدار F در ناحیه بحرانی قرار بگیرد نشان می‌دهد که تغییر ساختاری رخ داده است و ضرایب در این دو دوره متفاوت هستند.

آزمون علیت گرانجر در Stata

آزمون علیت گرانجر را می‌توان با استفاده از مدل‌های خودرگرسیون برداری (VAR) انجام داد اما چون این مدل‌ها در فصل ۱۸ بررسی می‌شوند لذا در اینجا از روش ساده‌تری استفاده می‌کنیم. فرض کنید که رابطه علین بین دو متغیر y1 و y2 را می‌خواهیم با پنج وقفه بررسی کنیم.
ابتدا بررسی می‌کنیم که آیا y2 علت y1 هست یا نه. بدین منظور معادله زیر را برآورد می‌کنیم:

```
reg  y1  l(1/5).y1  l(1/5).y2
```

حال آزمون صفر بودن ضرایب y2 را به فرمان زیر انجام می‌دهیم:

test (l1.y2=l2.y2=l3.y2=l4.y2=l5.y2=0)

نتیجه آزمون عبارت است از:

```
. test (l1.y2=l2.y2=l3.y2=l4.y2=l5.y2=0)

 ( 1)  L.y2 - L2.y2 = 0
 ( 2)  L.y2 - L3.y2 = 0
 ( 3)  L.y2 - L4.y2 = 0
 ( 4)  L.y2 - L5.y2 = 0
 ( 5)  L.y2 = 0

       F(  5,    16) =    0.98
            Prob > F =    0.4604
```

چون مقدار F=۰/۹۸ در ناحیه بحرانی قرار ندارد (زیرا احتمال بزرگتر از ۰/۰۵ است)، لذا y2علت y1 نیست.

مشابه بحث فوق بررسی می‌کنیم که آیا y1 علت y2 هست یا نه. بدین منظور معادله زیر را برآورد می‌کنیم:

reg y2 l(1/5).y1 l(1/5).y2

حال آزمون صفر بودن ضرایب y1 را به فرمان زیر انجام می‌دهیم:

test (l1.y1=l2.y1=l3.y1=l4.y1=l5.y1=0)

نتیجه آزمون عبارت است از:

```
. test (l1.y1=l2.y1=l3.y1=l4.y1=l5.y1=0)

 ( 1)  L.y1 - L2.y1 = 0
 ( 2)  L.y1 - L3.y1 = 0
 ( 3)  L.y1 - L4.y1 = 0
 ( 4)  L.y1 - L5.y1 = 0
 ( 5)  L.y1 = 0

       F(  5,    16) =    0.20
            Prob > F =    0.9566
```

چون مقدار F=۰/۲۰ در ناحیه بحرانی قرار ندارد (زیرا احتمال بزرگتر از ۰/۰۵ است)، لذا y1 علت y2 نیست.

ضمیمه

جداول آماری

مقادیر توزیع $t_{\alpha,n}$

n	0.25	0.1	0.05	0.025	0.01	0.005
1	1.000	3.078	6.314	12.706	31.821	63.657
2	.816	1.886	2.920	4.303	6.965	9.925
3	.765	1.638	2.353	3.182	4.541	5.841
4	.741	1.533	2.132	2.776	3.747	4.604
5	.727	1.476	2.015	2.571	3.365	4.032
6	.718	1.440	1.943	2.447	3.143	3.707
7	.711	1.415	1.895	2.365	2.998	3.499
8	.706	1.397	1.860	2.306	2.896	3.355
9	.703	1.383	1.833	2.262	2.821	3.250
10	.700	1.372	1.812	2.228	2.764	3.169
11	.697	1.363	1.796	2.201	2.718	3.106
12	.695	1.356	1.782	2.179	2.681	3.055
13	.694	1.350	1.771	2.160	2.650	3.012
14	.692	1.345	1.761	2.145	2.624	2.977
15	.691	1.341	1.753	2.131	2.602	2.947
16	.690	1.337	1.746	2.120	2.583	2.921
17	.689	1.333	1.740	2.110	2.567	2.898
18	.688	1.330	1.734	2.101	2.552	2.878
19	.688	1.328	1.729	2.093	2.539	2.861
20	.687	1.325	1.725	2.086	2.528	2.845
21	.686	1.323	1.721	2.080	2.518	2.831
22	.686	1.321	1.717	2.074	2.508	2.819
23	.685	1.319	1.714	2.069	2.500	2.807
24	.685	1.318	1.711	2.064	2.492	2.797
25	.684	1.316	1.708	2.060	2.485	2.787
26	.684	1.315	1.706	2.056	2.479	2.779
27	.684	1.314	1.703	2.052	2.473	2.771
28	.683	1.313	1.701	2.048	2.467	2.763
29	.683	1.311	1.699	2.045	2.462	2.756
30	.683	1.310	1.697	2.042	2.457	2.750
35	.682	1.306	1.690	2.030	2.438	2.724
40	.681	1.303	1.684	2.021	2.423	2.704
45	.680	1.301	1.679	2.014	2.412	2.690
50	.679	1.299	1.676	2.009	2.403	2.678
60	.679	1.296	1.671	2.000	2.390	2.660
70	.678	1.294	1.667	1.994	2.381	2.648
80	.678	1.292	1.664	1.990	2.374	2.639
90	.677	1.291	1.662	1.987	2.368	2.632
100	.677	1.290	1.660	1.984	2.364	2.626
∞	.674	1.282	1.645	1.960	2.326	2.576

مقادیر توزیع $\chi^2_{\alpha,n}$

n	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	.00004	.0002	.001	.004	.02	.10	.45	1.32	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49

مقادیر توزیع $F_{0.05,n_1,n_2}$

درجه آزادی صورت $= n_1$

n_2	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

n_2	10	12	15	20	30	40	50	60	∞
1	241.88	243.91	245.95	248.01	250.10	251.14	252.20	252.20	254.19
2	19.40	19.41	19.43	19.45	19.46	19.47	19.48	19.48	19.49
3	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.57	8.53
4	5.96	5.91	5.86	5.80	5.75	5.72	5.69	5.69	5.63
5	4.74	4.68	4.62	4.56	4.50	4.46	4.43	4.43	4.37
6	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.74	3.67
7	3.64	3.57	3.51	3.44	3.38	3.34	3.30	3.30	3.23
8	3.35	3.28	3.22	3.15	3.08	3.04	3.01	3.01	2.93
9	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.79	2.71
10	2.98	2.91	2.85	2.77	2.70	2.66	2.62	2.62	2.54
15	2.54	2.48	2.40	2.33	2.25	2.20	2.16	2.16	2.07
20	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.95	1.85
25	2.24	2.16	2.09	2.01	1.92	1.87	1.82	1.82	1.72
30	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.74	1.63
40	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.64	1.52
50	2.03	1.95	1.87	1.78	1.69	1.63	1.58	1.58	1.45
70	1.97	1.89	1.81	1.72	1.62	1.57	1.50	1.50	1.36
100	1.93	1.85	1.77	1.68	1.57	1.52	1.45	1.45	1.30
∞	1.83	1.75	1.67	1.57	1.46	1.39	1.34	1.31	1.30

مقادیر توزیع $F_{.01, n_1, n_2}$

درجه آزادی صورت = n_1

n_2	1	2	3	4	5	6	7	8	9
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
∞	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43

n_2	10	12	15	20	30	40	50	60	∞
1	6055.85	6106.32	6157.28	6208.73	6260.65	6286.78	6313.03	6313.03	6362.68
2	99.40	99.42	99.43	99.45	99.47	99.47	99.48	99.48	99.50
3	27.23	27.05	26.87	26.69	26.50	26.41	26.32	26.32	26.14
4	14.55	14.37	14.20	14.02	13.84	13.75	13.65	13.65	13.47
5	10.05	9.89	9.72	9.55	9.38	9.29	9.20	9.20	9.03
6	7.87	7.72	7.56	7.40	7.23	7.14	7.06	7.06	6.89
7	6.62	6.47	6.31	6.16	5.99	5.91	5.82	5.82	5.66
8	5.81	5.67	5.52	5.36	5.20	5.12	5.03	5.03	4.87
9	5.26	5.11	4.96	4.81	4.65	4.57	4.48	4.48	4.32
10	4.85	4.71	4.56	4.41	4.25	4.17	4.08	4.08	3.92
15	3.80	3.67	3.52	3.37	3.21	3.13	3.05	3.05	2.88
20	3.37	3.23	3.09	2.94	2.78	2.69	2.61	2.61	2.43
25	3.13	2.99	2.85	2.70	2.54	2.45	2.36	2.36	2.18
30	2.98	2.84	2.70	2.55	2.39	2.30	2.21	2.21	2.02
40	2.80	2.66	2.52	2.37	2.20	2.11	2.02	2.02	1.82
50	2.70	2.56	2.42	2.27	2.10	2.01	1.91	1.91	1.70
70	2.59	2.45	2.31	2.15	1.98	1.89	1.78	1.78	1.56
100	2.50	2.37	2.22	2.07	1.89	1.80	1.69	1.69	1.45
∞	2.34	2.20	2.06	1.90	1.72	1.61	1.50	1.50	1.16

مقادیر آماره دوربین- واتسن (DW)

	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$		$k=10$		$k=15$	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21				
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15	.16	3.30		
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10	.20	3.18		
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06	.24	3.07		
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02	.29	2.97		
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99	.34	2.89	.06	3.68
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96	.38	2.81	.09	3.58
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94	.42	2.73	.12	3.55
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92	.47	2.67	.15	3.41
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90	.51	2.61	.19	3.33
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89	.54	2.57	.22	3.25
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88	.58	2.51	.26	3.18
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86	.62	2.47	.29	3.11
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85	.65	2.43	.33	3.05
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84	.68	2.40	.36	2.99
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83	.71	2.36	.39	2.94
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83	.74	2.33	.43	2.99
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82	.77	2.31	.46	2.84
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81	.80	2.28	.49	2.80
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81	.82	2.26	.52	2.75
35	1.40	1.52	1.34	1.53	1.28	1.65	1.22	1.73	1.16	1.80	.85	2.24	.55	2.72
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80	.87	2.22	.58	2.68
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80	.89	2.20	.60	2.65
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79	.91	2.18	.63	2.61
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79	.93	2.16	.65	2.59
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79	.95	2.15	.68	2.56
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78	1.04	2.09	.79	2.44
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77	1.11	2.04	.88	2.35
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77	1.17	2.01	.96	2.28
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77	1.22	1.98	1.03	2.23
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77	1.27	1.96	1.09	2.18
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77	1.30	1.95	1.14	2.15
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77	1.34	1.94	1.18	2.12
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77	1.37	1.93	1.22	2.09
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77	1.40	1.92	1.26	2.07
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78	1.42	1.91	1.29	2.06
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78	1.44	1.90	1.32	2.04
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78	1.46	1.90	1.35	2.03

تعداد ضرایب به استثنای عرض از مبدأ=k

تعداد مشاهدات=n